GeoDa

An Introduction to Spatial Data Analysis

Luc Anselin, PhD
Julia Koschinsky
University of Illinois, Urbana-Champaign

Spatial Analysis Laboratory
http://sal.uiuc.edu

The GeoDa Team

- **Luc Anselin**: Overall Direction

- **Ibnu Syabri**: EDA and ESDA, Flows

- **Youngihn Kho**: 3D Visualization, Graphics

- **Yun-Tien Lee**: Cross-Platform, wxWidgets

- **Julia Koschinsky**: Documentation and Training

Acknowledgments

- National Science Foundation
  - Center for Spatially Integrated Social Science
- National Cancer Institute
- Association of Teachers of Preventive Medicine / Centers for Disease Control

Outline

- Why – Background

- How – Functionality

- What is Next

Background

CSISS

- Center for Spatially Integrated Social Science
  - NSF Infrastructure Project 1999-2004
  - headquartered UC Santa Barbara
  - Software Tools Program at UIUC
- Mission
  - promote spatial thinking and spatial analysis in social science

http://www.csiss.org/

http://www.csiss.org/

Q▾ Google

Center for Spatially Integrated Social Science
Spatial Resources for the Social Sciences

The CSISS Mission recognizes the growing significance of space, spatiality, location, and place in social science research. It seeks to develop unrestricted access to tools and perspectives that will advance the spatial analytic capabilities of researchers throughout the social sciences. CSISS is funded by the National Science Foundation under its program of support for infrastructure in the social and behavioral sciences.

CSISS News   Upgrade to GeoDa™ 0.9.5-i, DynESDA Replacement

Core Programs	**Learning Resources**	**Spatial Resources**	**Spatial Tools**
CSISS has six research initiatives and a professional development program for undergraduate instructors.	These introductory materials include **CSISS Classics** and **select video clips** from the CSISS summer workshops.	CSISS has compiled e-journals, bibliographies, and other spatial resources for the social sciences.	**Spatial Tools Search Engine** **Select Tools** **Links to Portals** **GeoDa™** **Tobler's Flow Mapper**

Search Engines	**CSISS Events**	**Community Center**	
Try out one of our custom search engines to find spatial analysis resources on the Internet.	Here's where you'll find information and registration for workshops, conferences and specialist meetings.	Join the forums, or if your organization relates to our mission and goals, register as a CSISS affiliate.	personnel, and sitemap. Our Strategic Plan and Annual Reports are also found here.

Core Programs | Learning Resources | Spatial Resources | Spatial Tools | Search Engines | CSISS Events | Community Center | About CSISS
Site Map | Site Search | Contact CSISS | Plug-ins | Privacy Policy | Site Credits | Home

CSISS Tools Project

- Goals
  - facilitate dissemination of spatial analysis software tools to social scientists
  - develop a library of spatial data analysis modules
  - develop prototypes implementing state of the art methods
  - initiate and nurture a community
- Products
  - GeoDa
  - PySAL and PySpace

http://www.csiss.org/clearinghouse/

| Return Home | Core Programs | Learning Resources | Spatial Resources | Spatial Tools | Search Engines | CSISS Events | Community Center | About CSISS |

Center for Spatially Integrated Social Science

back

Site Search

CSISS Tools Clearinghouse

The **CSISS Tools Clearinghouse** is intended to grow into a robust collection of spatial analysis software, software links, and links to information about tools for spatial analysis. The development of these tools is a lively research area and the goal of this clearinghouse is to provide up-to-date information on available tools. The clearinghouse is comprised of:

○ **Search Engine**
Search a continuously updated, comprehensive index of the CSISS Select Tools and Links to Portals.

○ **Select Tools**
Browse through tools particularly suited to the analysis of spatial phenomena.

○ **Portal Links**
A listing of useful collections of software tools for anyone interested in Spatial Analysis, or those looking for specific tools.

○ **CSISS Tools** (offsite)
The home of the software tools development efforts under CSISS, carried out in the Spatial Analysis Laboratory of the Department of Agricultural and Consumer Economics at the University of Illinois, Urbana-Champaign.

New - GeoDa 0.9.5-i, beta release software for ESDA with dynamically linked windows.
New - R-Geo, a developing effort to promote spatial data analysis software in the R language.

○ **Tobler's Flow Mapper**
New - Software for interactive flow mapping.

Your help is requested in suggesting tools, collections of tools, other portals, and methods that should be represented in this collection - please send these to the Tools Manager, Luc Anselin at anselin@uiuc.edu.

Core Programs | Learning Resources | Spatial Resources | Spatial Tools | Search Engines | CSISS Events | Community Center | About CSISS
Site Map | Contact CSISS | Plug-ins | Privacy Policy | Site Credits | Home

Copyright © 2001-2004 by Regents of University of California, Santa Barbara
Tools Clearinghouse: Kevin Konty
Webmaster: Gamaiel Zavala
Search Engines: Eric White

Spatial Analysis and GIS

- Early Efforts (1980s)
  - linking statistical packages and GIS
  - little in terms of spatial statistics
- Linked Spatial Stats and GIS
  - lots of examples in 1990s
    - Arcinfo and SpaceStat, ArcView-DynESDA, ArcView-XGobi, SAGE, using MapObjects, etc.
    - geovisualization: cdv, GeoVISTA, ...
  - limited functionality, often inflexible architecture
  - slow, limited capacity for larger data sets

GeoDa Requirements

- Free Standing
  - no GIS
- Modular
  - allow extensibility
- Fast
  - true interactive data exploration, linking and brushing
- Large Data Analysis
  - 100,000+

GeoDa Objectives

- An **Introduction** to Spatial Data Analysis
  - visual, interactive, user-friendly
  - path from geo-visualization, EDA and ESDA to spatial regression
  - aimed at non (or not yet) GIS users
- Free Download
- Support through OpenSpace Mailing List

http://sal.uiuc.edu/mailman/listinfo/openspace

http://geoda.uiuc.edu

GeoDa User Base

Bar chart showing GeoDa user base growth: Feb 2003: 1; Nov 03: 912; May 04: 2,070; Dec 04: 4,025; July 05: 7,249
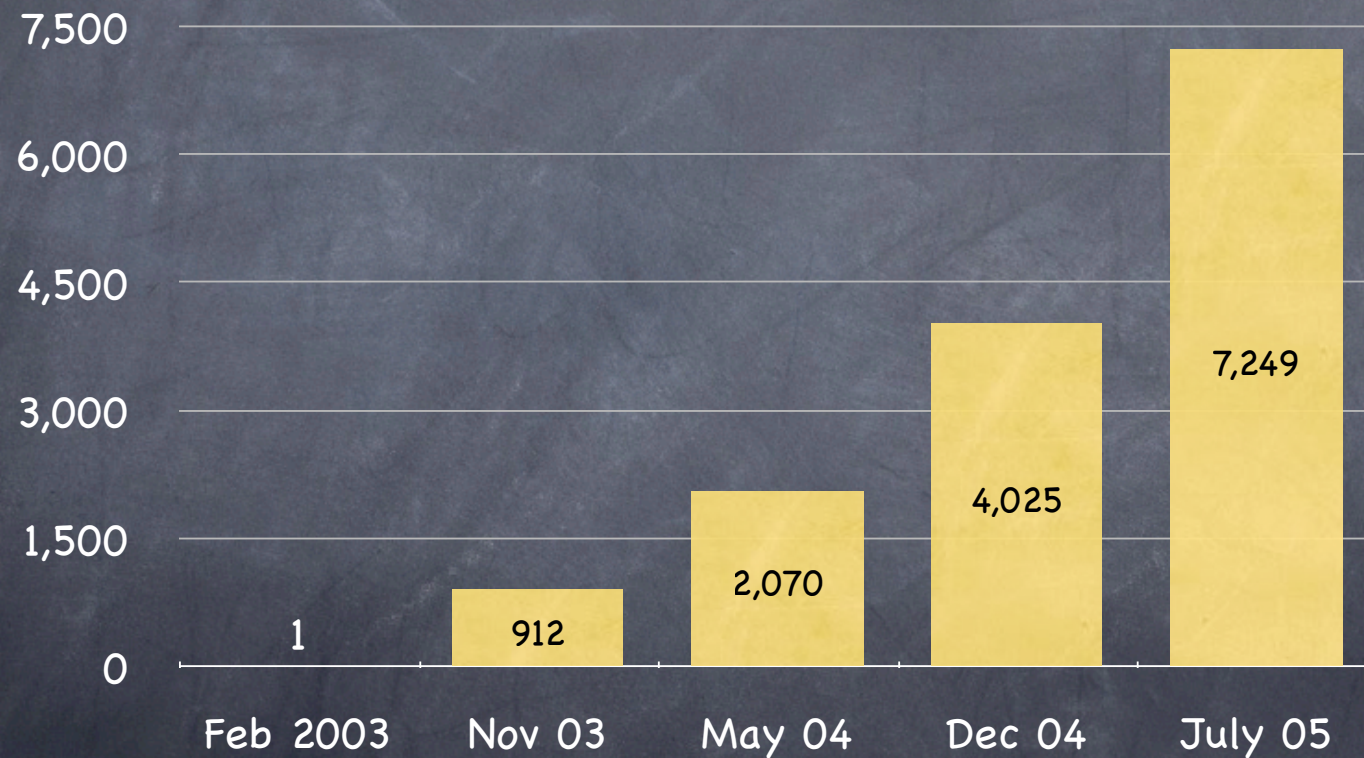
Geographical Distribution

GeoDa Users
- 0
- 1 - 11
- 12 - 17
- 18 - 39
- 40 - 95
- 96 - 747

May 14, 2004

Produced by Spatial Analysis Lab (http://sal.agecon.uiuc.edu).
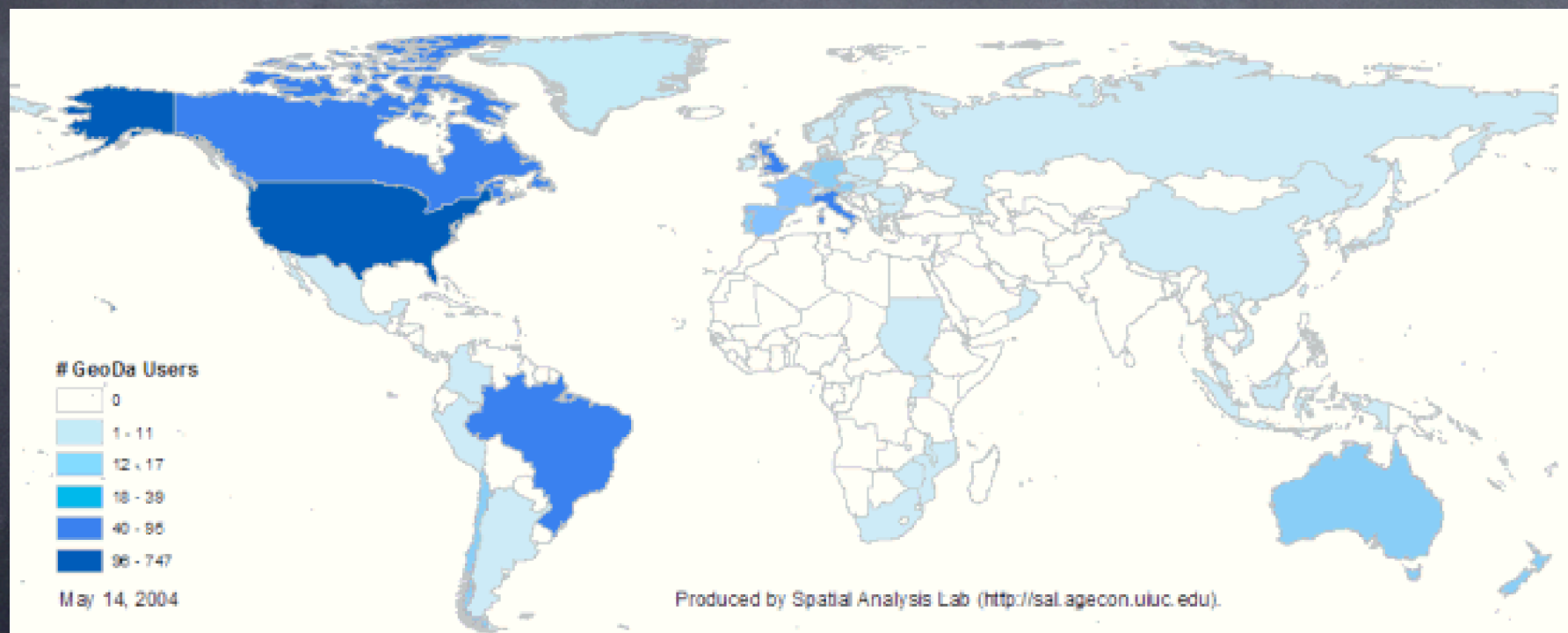
GeoDa & Crime Analysis

What GeoDa is designed for

- Analysis of lattice data: Areas & fixed points

    - Events (e.g., address-level crimes) have to be aggregated to areas

    - Estimating underlying risk of crime: Smoothing

What GeoDa is <u>not</u> designed for

- Point pattern analysis

  - Event data: Location subject to uncertainty

- Geostatistical analysis

  - Sample points

Functionality
Principles

Main Toolbar
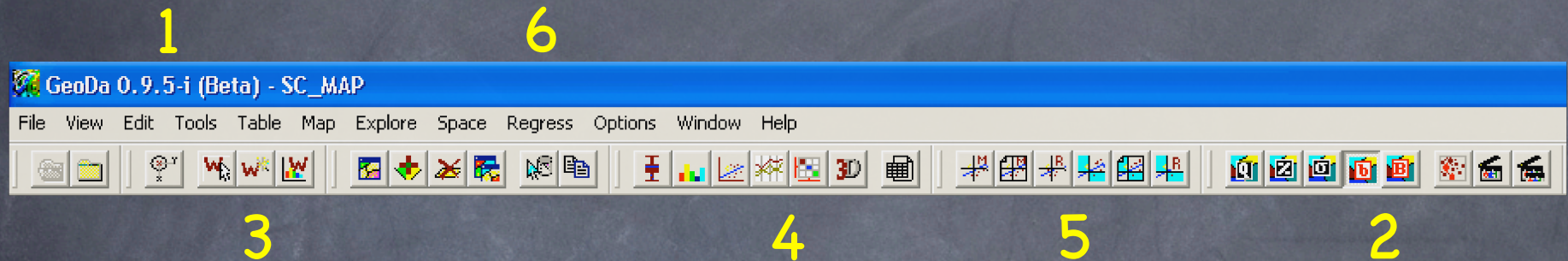
- Data Manipulation (1)
- Mapping (2)
- Spatial Weights (3)
- Multivariate (non-spatial) EDA (4)
- Spatial Autocorrelation (5)
- Spatial Regression (6)

Linking and Brushing

- Dynamic Graphics
  - user interaction with multiple statistical graphs and maps
- Linking
  - selection in one graph is simultaneously selected in all graphs
- Brushing
  - changing the selection is dynamically updated in all graphs and maps

Linking
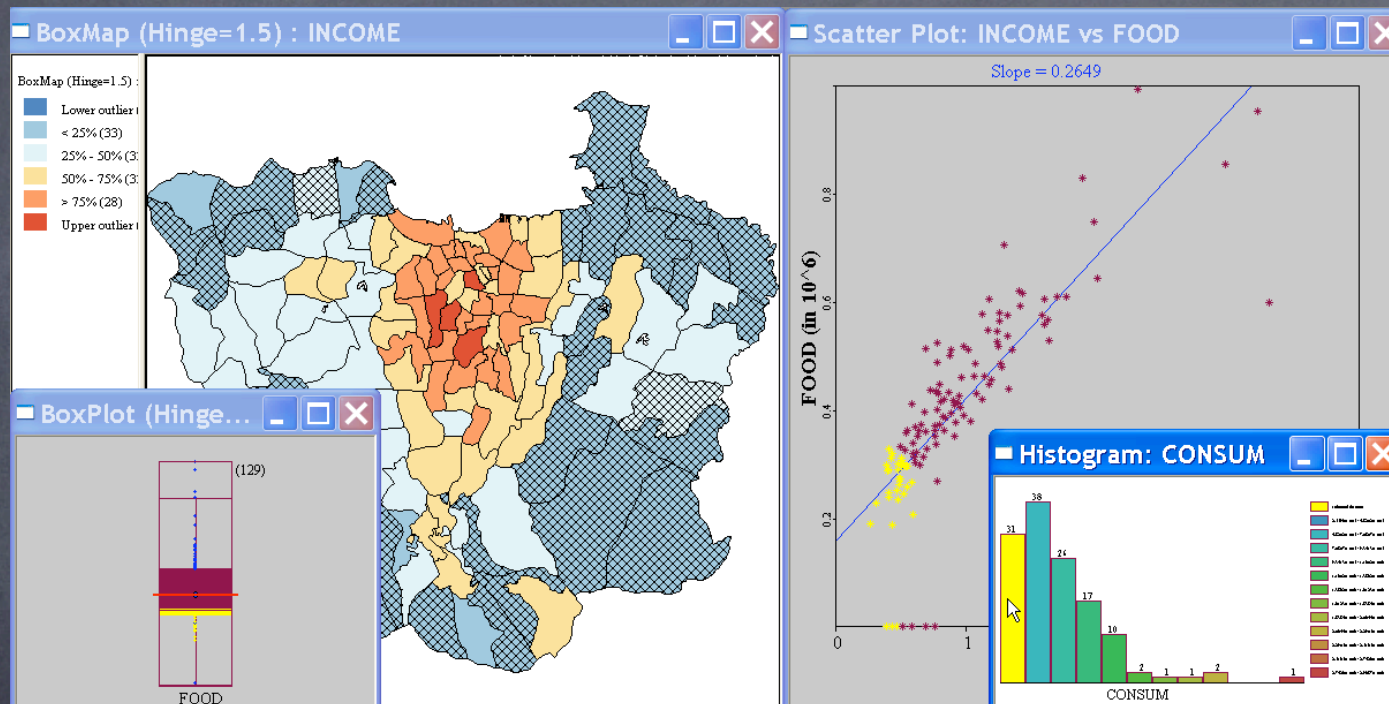
the selected observations in any graph are simultaneously highlighted in all other graphs
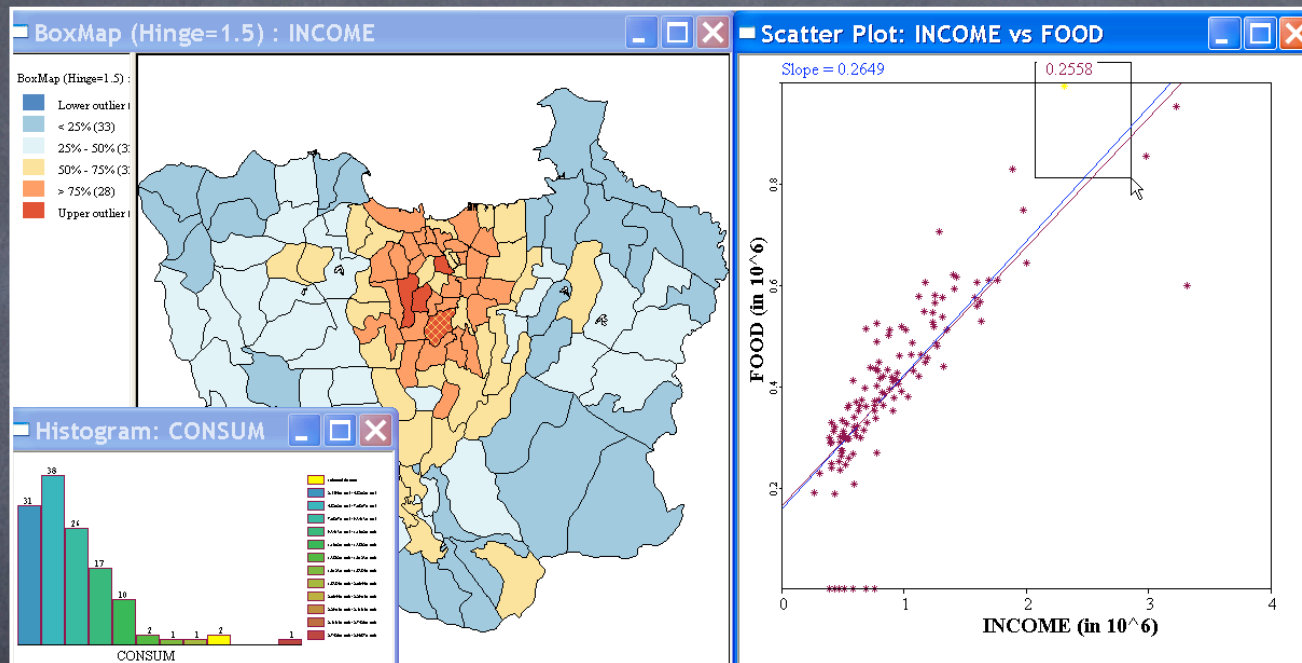
Brushing a Scatter Plot

- changing the selection in the scatter plot is simultaneously highlighted in all other graphs

Brushing a Scatter Plot

- changing the selection in the scatter plot is simultaneously highlighted in all other graphs

Brushing a Scatter Plot

- changing the selection in the scatter plot is simultaneously highlighted in all other graphs

Brushing a Map

changing the selection in the map is **simultaneously highlighted** in all other graphs

Brushing a Map

- changing the selection in the map is simultaneously highlighted in all other graphs

Brushing a Map
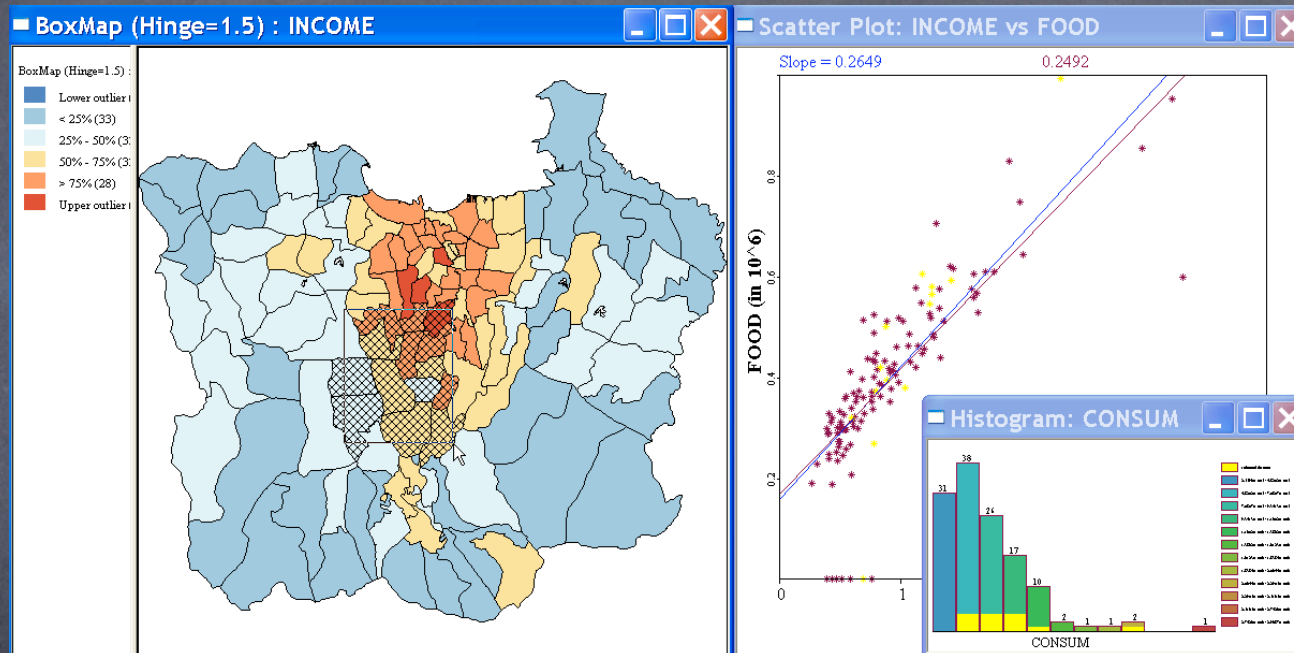
- changing the selection in the map is simultaneously highlighted in all other graphs

Functionality

Overview

Data Manipulation

- Input
  - polygon and point shape files
  - ascii point and boundary files
- Conversion/Creation
  - points to polygons (Thiessen polygons)
  - polygons to points (centroids)
  - regular grids
- Calculation
  - table calculator, rates, smoothing, queries
- Export
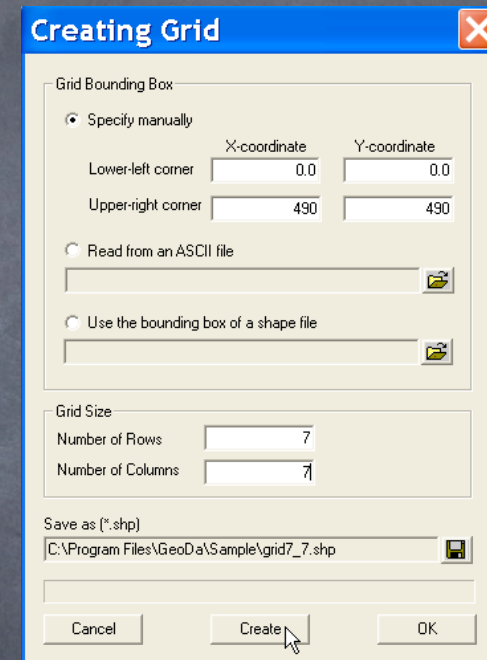 - boundary files, data sets, bit maps
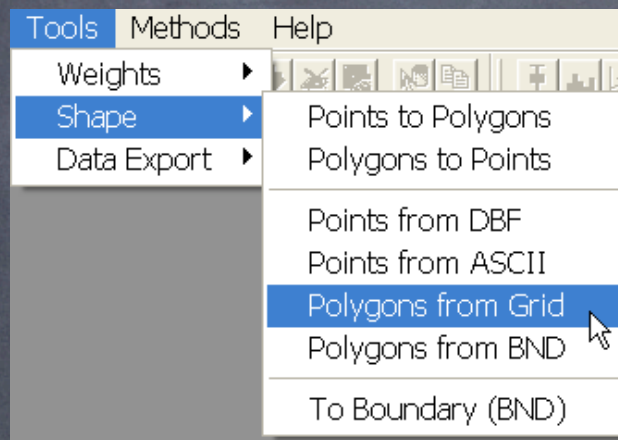
Thiessen Polygons
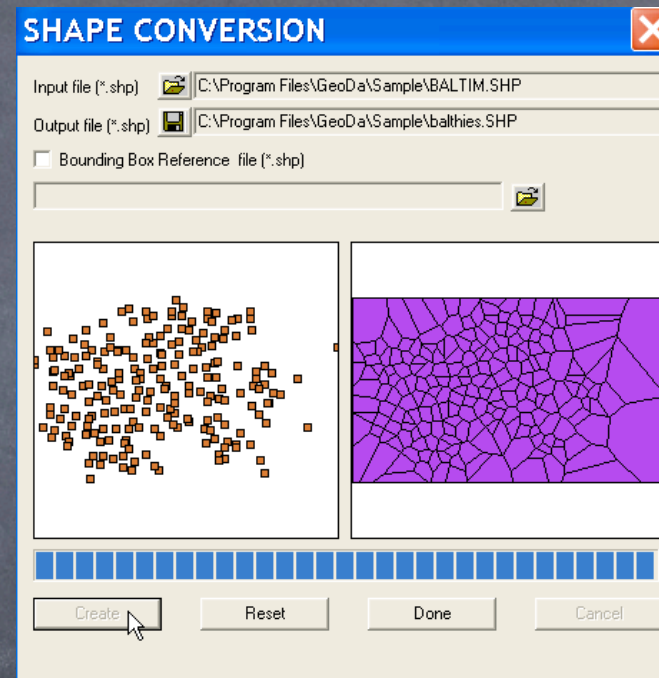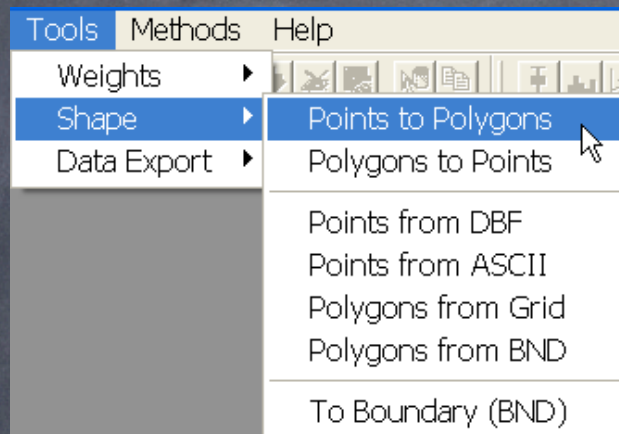
@ all points in the polygon are closer to the center point than to any other point = market area

Create Dummy Variable
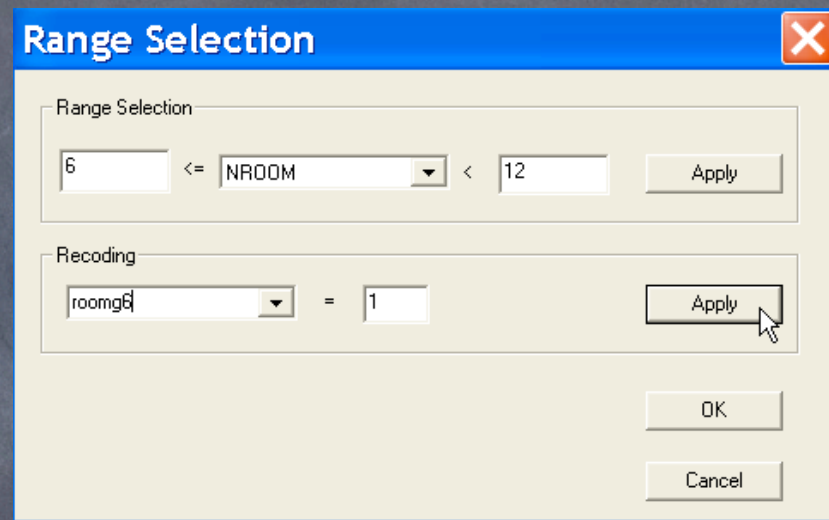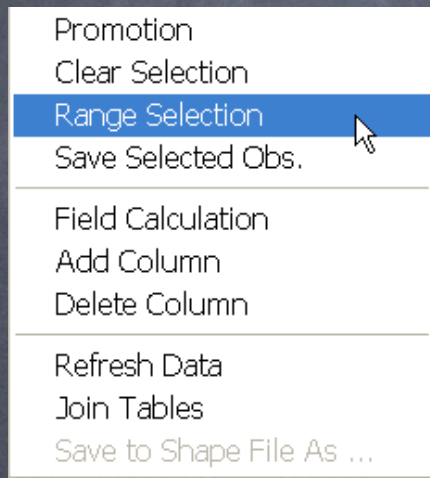
use Table Calculator to create indicator variables that meet specified criteria: both in table and from map (save selected)

Regular Lattices

create regular grid by specifying bounding box and number of grid cells

demonstration

Mapping

- Choropleth Maps
  - quantile, percentile, standard deviational
- Outlier Maps
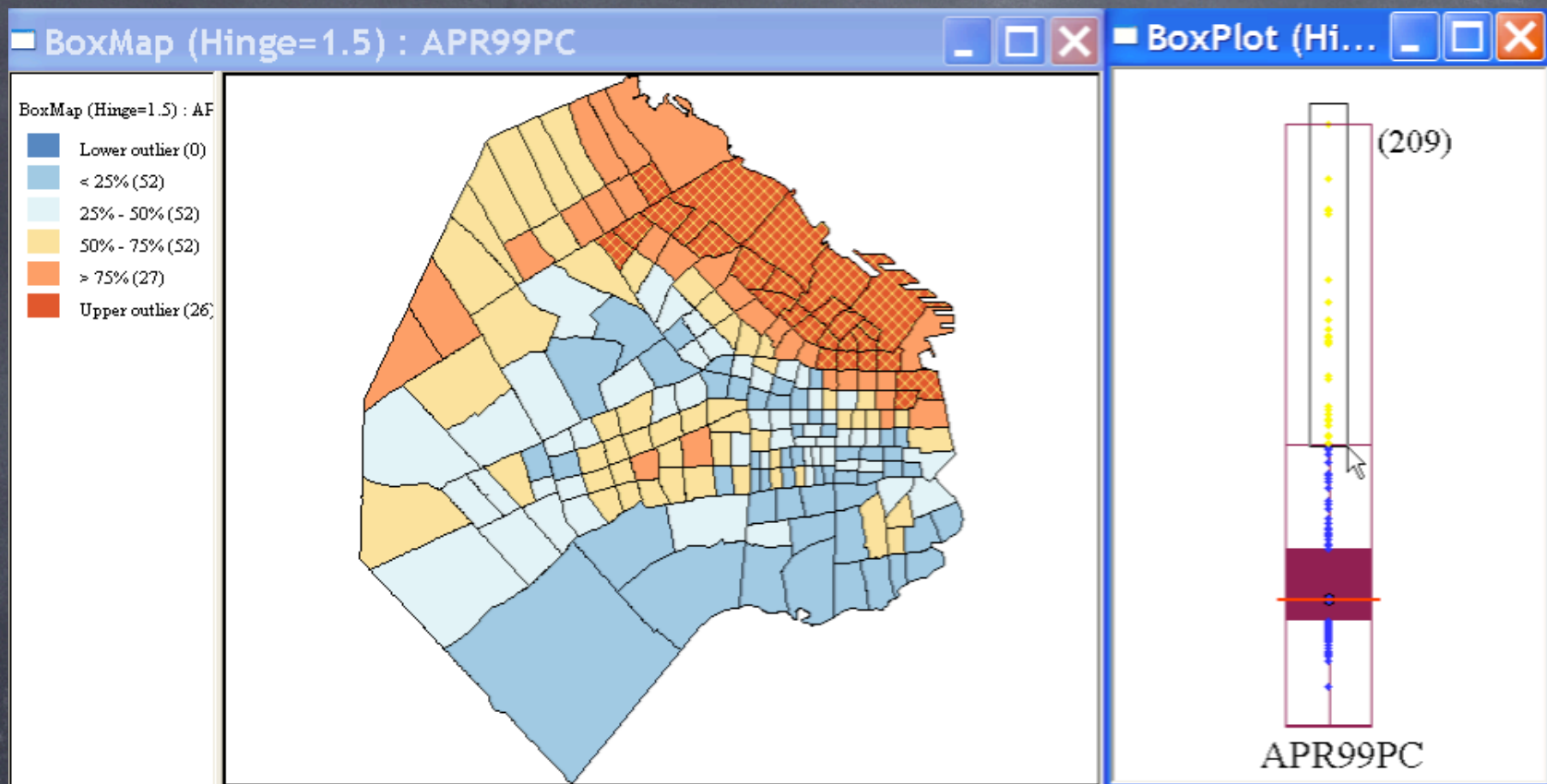 - box map
- Rate Smoothing
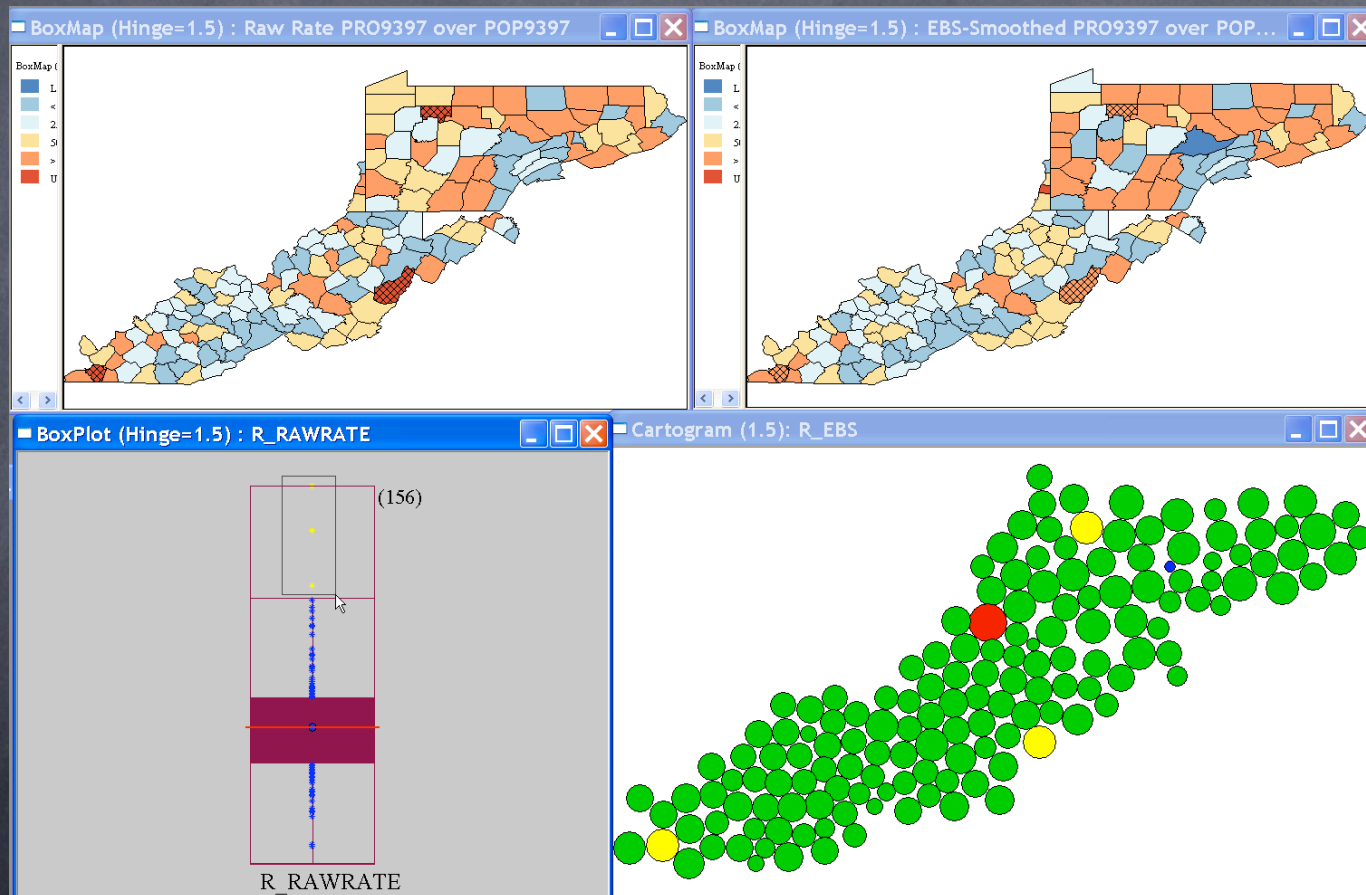 - Empirical Bayes
  - spatial smoothing
- Cartogram
- Map Movie

Box Map with Box Plot
Buenos Aires Election Results
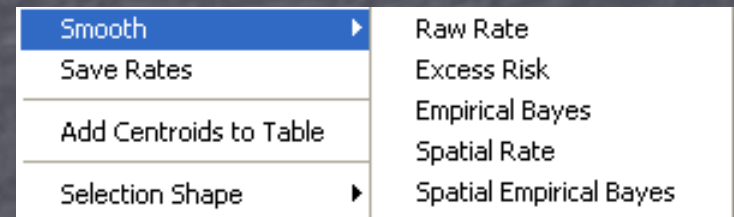
Smoothed Rates and Cartogram

Rate Smoothing

- Raw rates used to estimate underlying disease risk

- Differences in population size related to problem of variance instability and spurious outliers

- Rate smoothing addresses variance instability by borrowing strength from other spatial units

GeoDa Smoothing Options

GeoDa contains an excess risk map option and three smoothers:

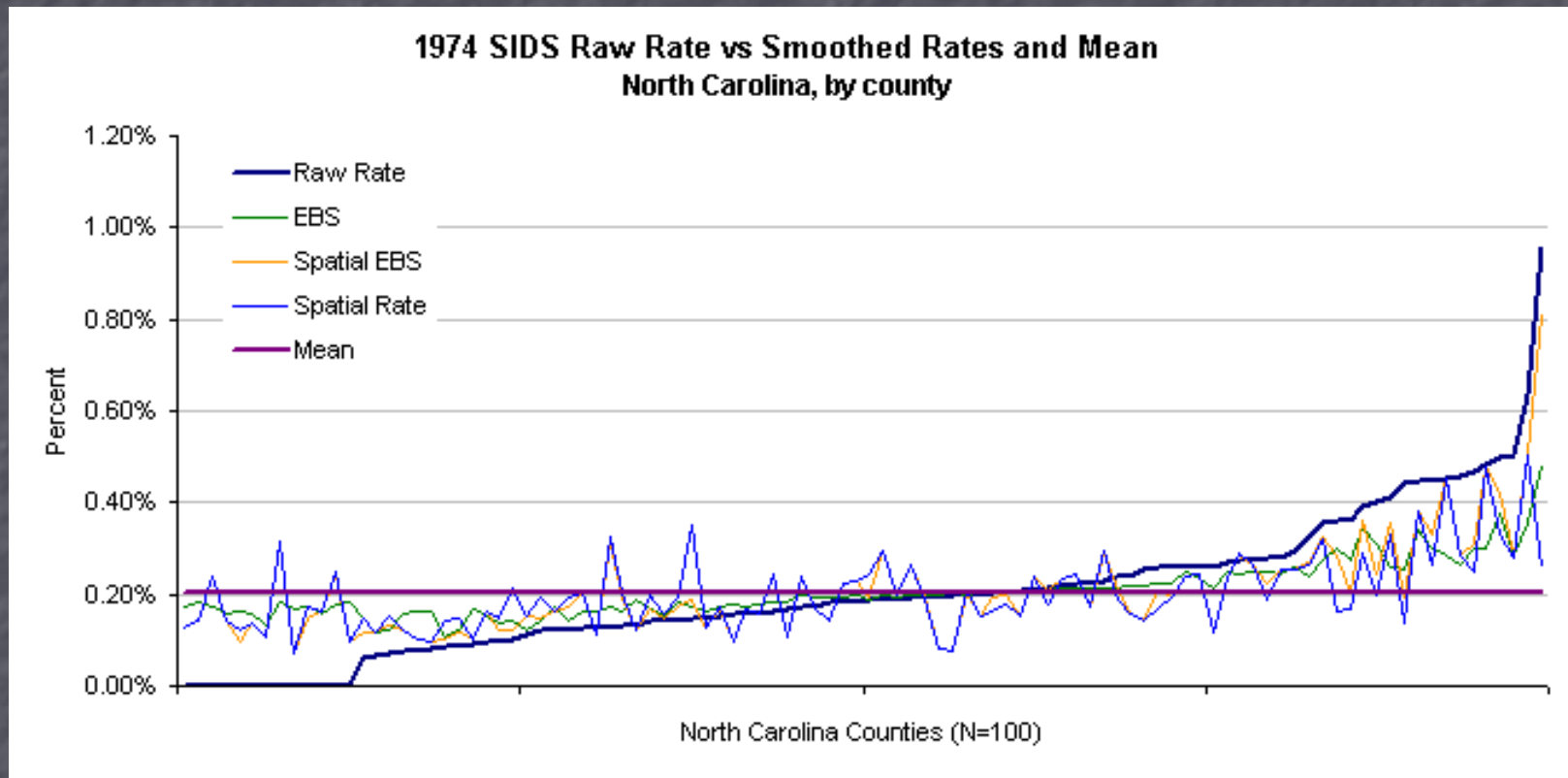| Smooth	▶	Raw Rate
Save Rates		Excess Risk
		Empirical Bayes
Add Centroids to Table		Spatial Rate
		Spatial Empirical Bayes
Selection Shape	▶	

Excess Risk: Expected risk based on
product of raw rate and average overall risk of all observations

Where Smoothers Borrow Strength From:
- **Empirical Bayes (EBS)**: Overall mean of the underlying risk distribution of all observations

- **Spatial Rate**: Neighbors, as defined in spatial weights matrix

- **Spatial Empirical Bayes**: Same as EBS but strength not borrowed from all observations, only regional subset
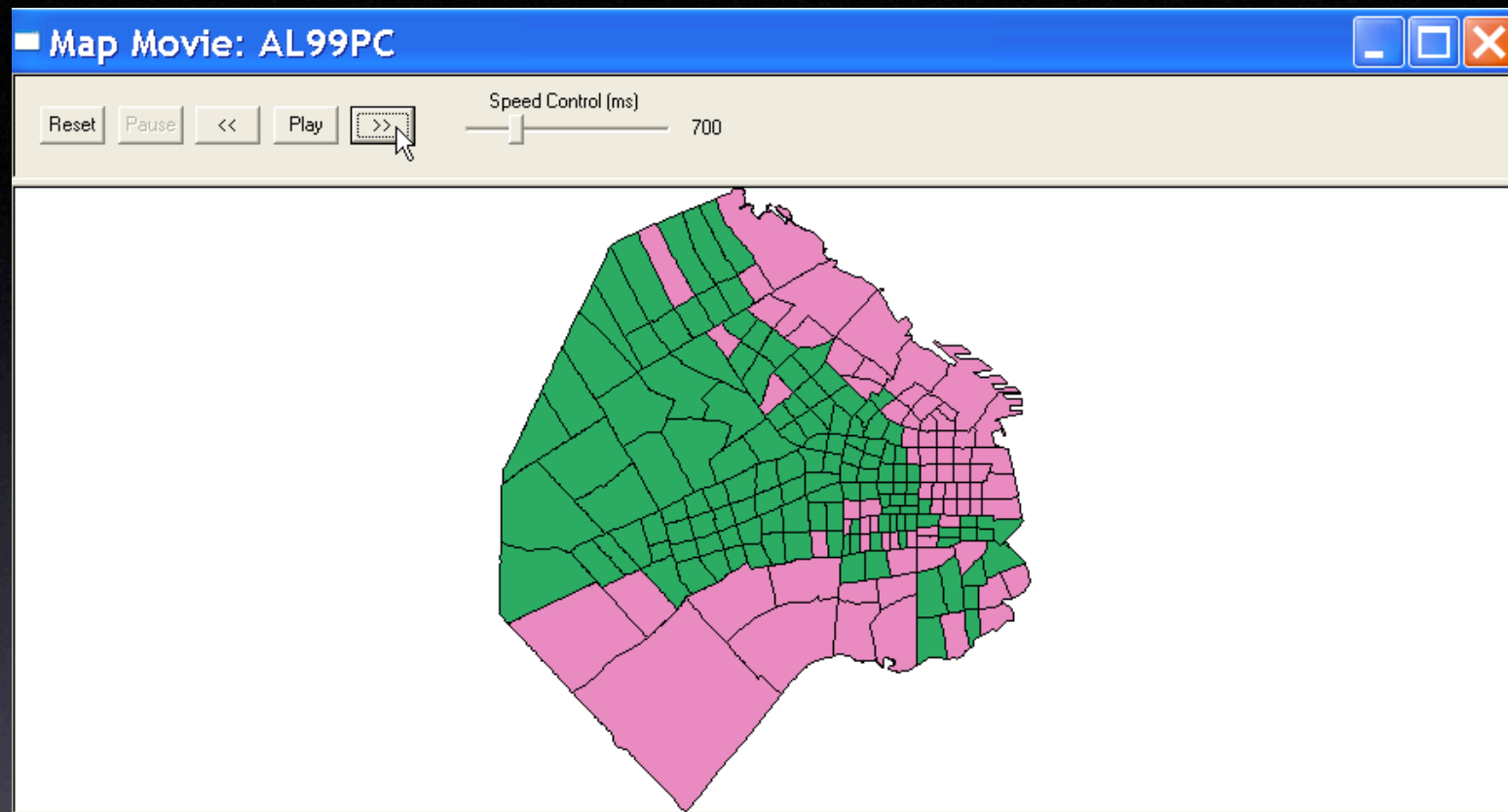
Rate Smoothing: Comparison
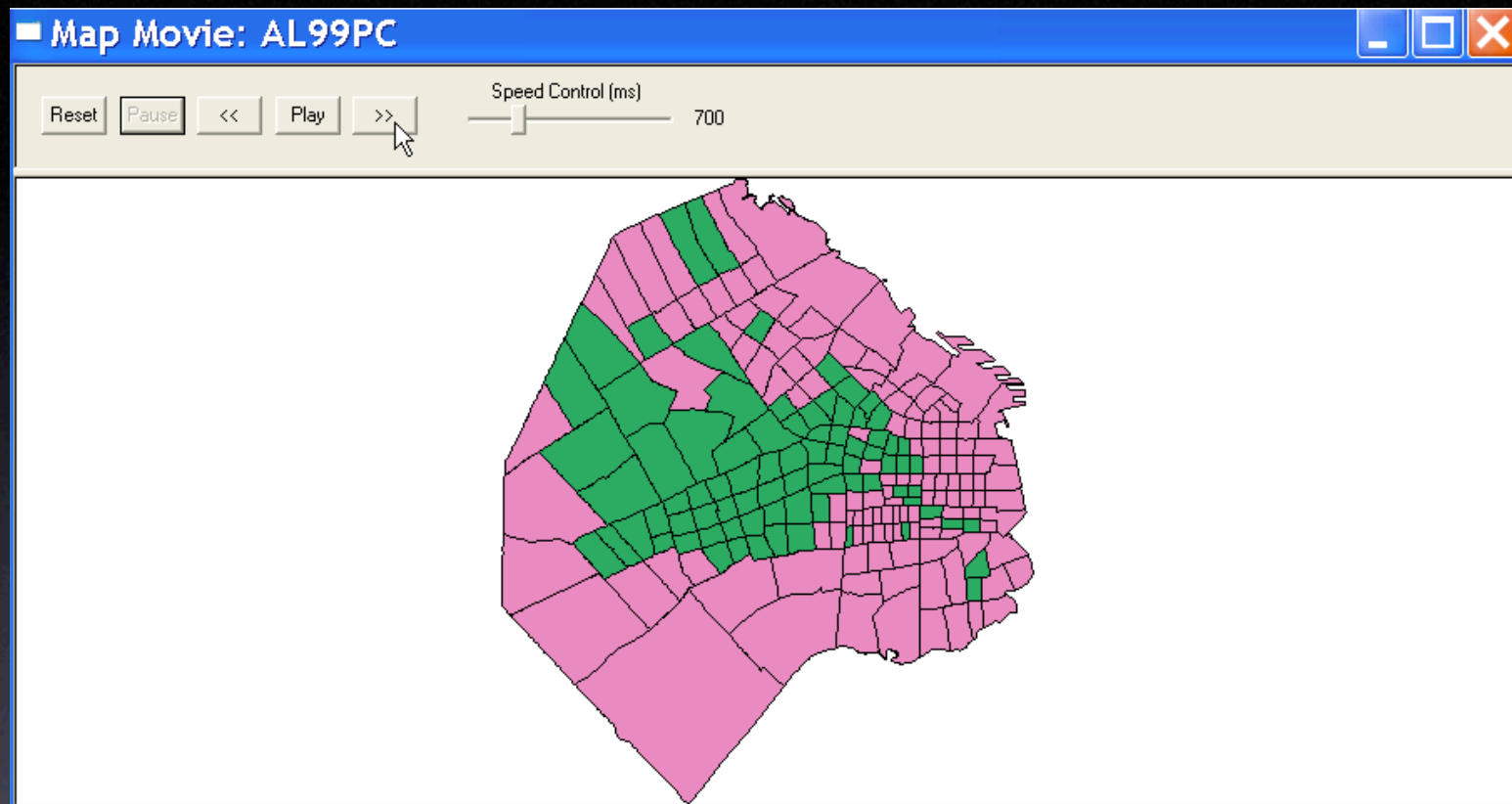

Raw rate compared to smoothed rates and mean
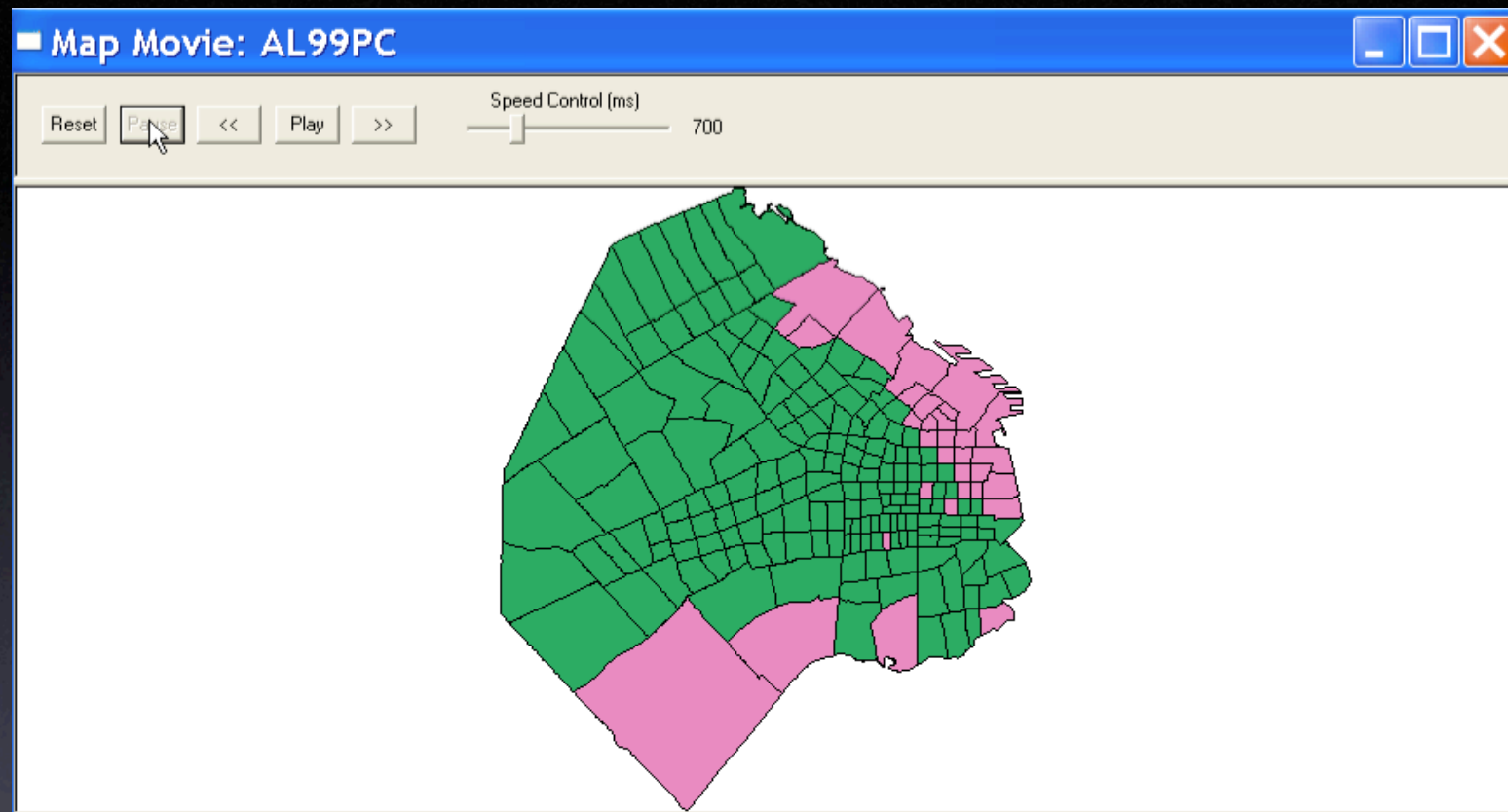
Map Movie
Buenos Aires Election Results

Map Movie
Buenos Aires Election Results

Map Movie
Buenos Aires Election Results

demonstration

Spatial Weights

- Contiguity Based

  - rook and queen

  - higher order contiguity

- Distance Based

  - distance band, k-nearest neighbors

- Weights Characteristics

  - connectivity, islands

Weights Dialogs

Islands

demonstration

Multivariate EDA

- Scatter Plot Matrix

- Parallel Coordinate Plot (PCP)

- Conditional Plots

    - conditional maps

    - conditional histograms

    - conditional box plots

    - conditional scatter plots

- 3D Scatter Plot
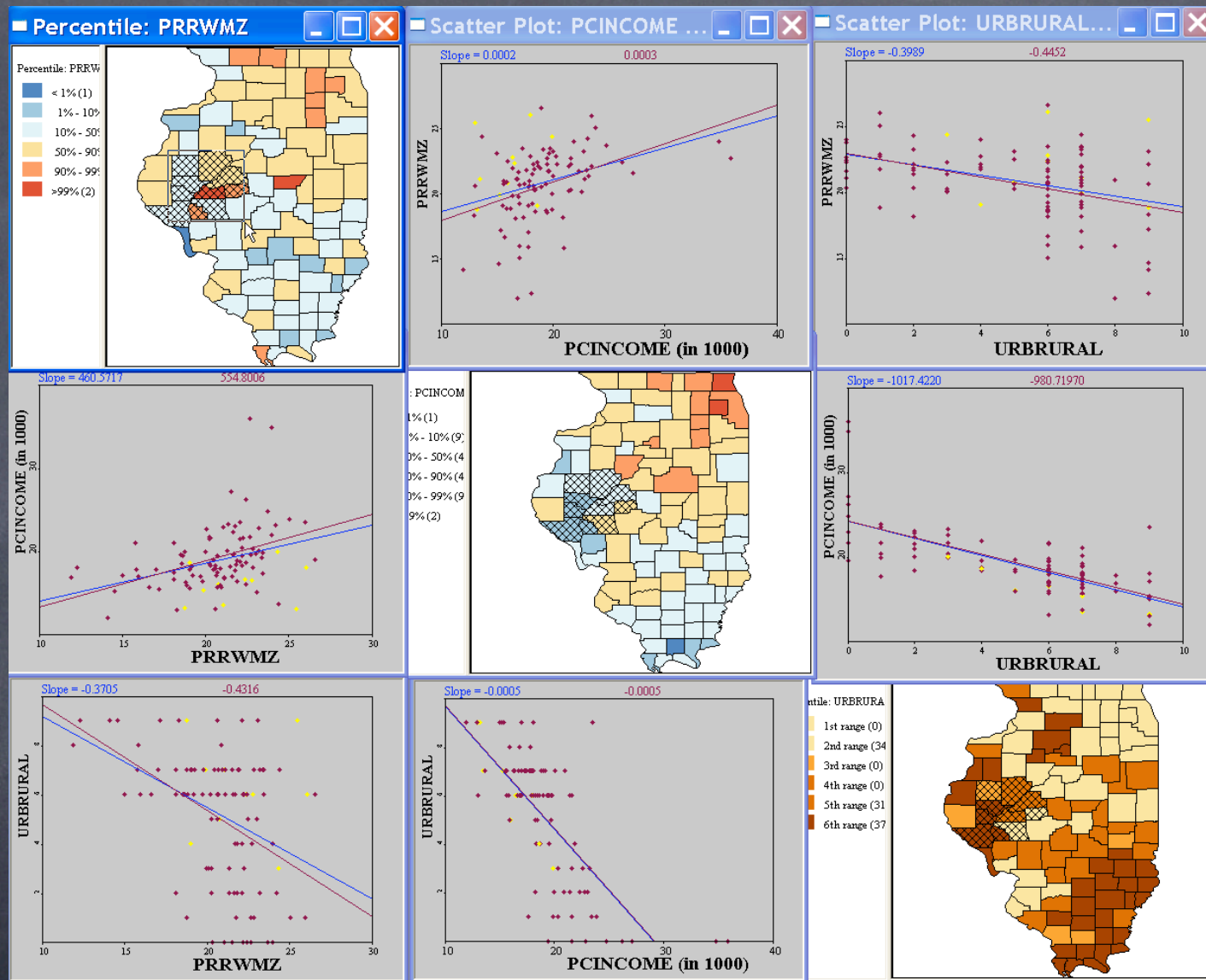
brushing the scatter plot matrix
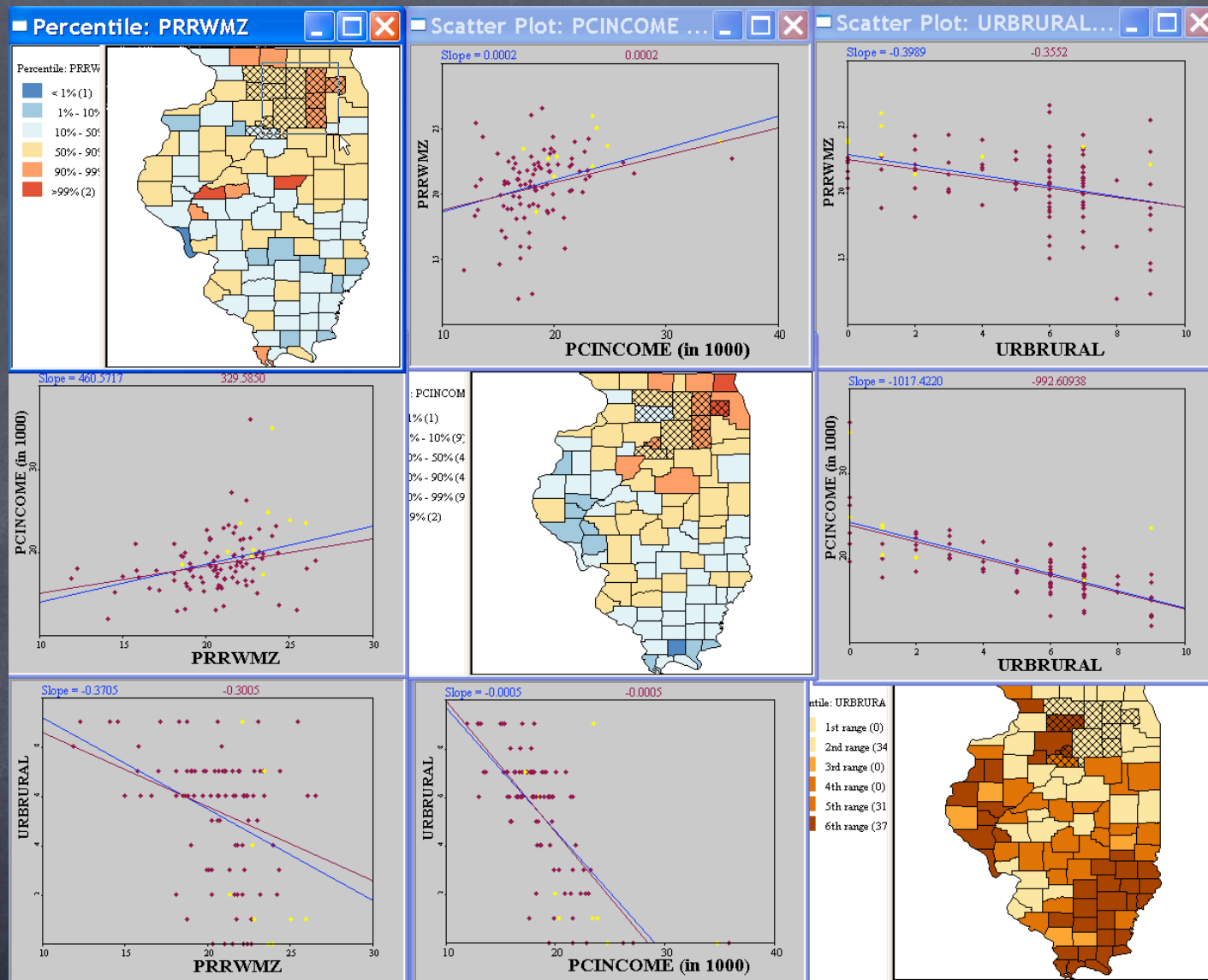
brushing the scatter plot matrix
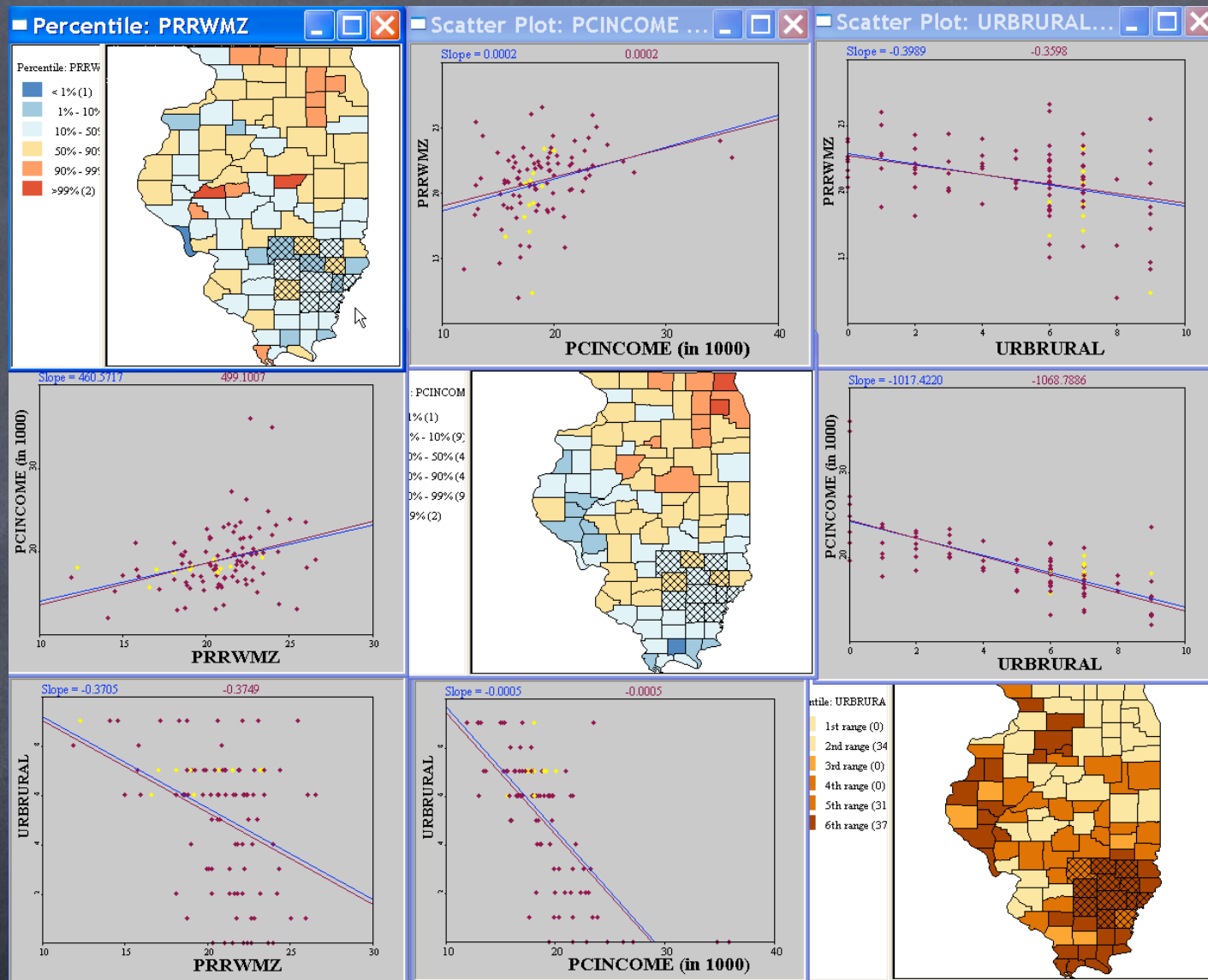
brushing the scatter plot matrix
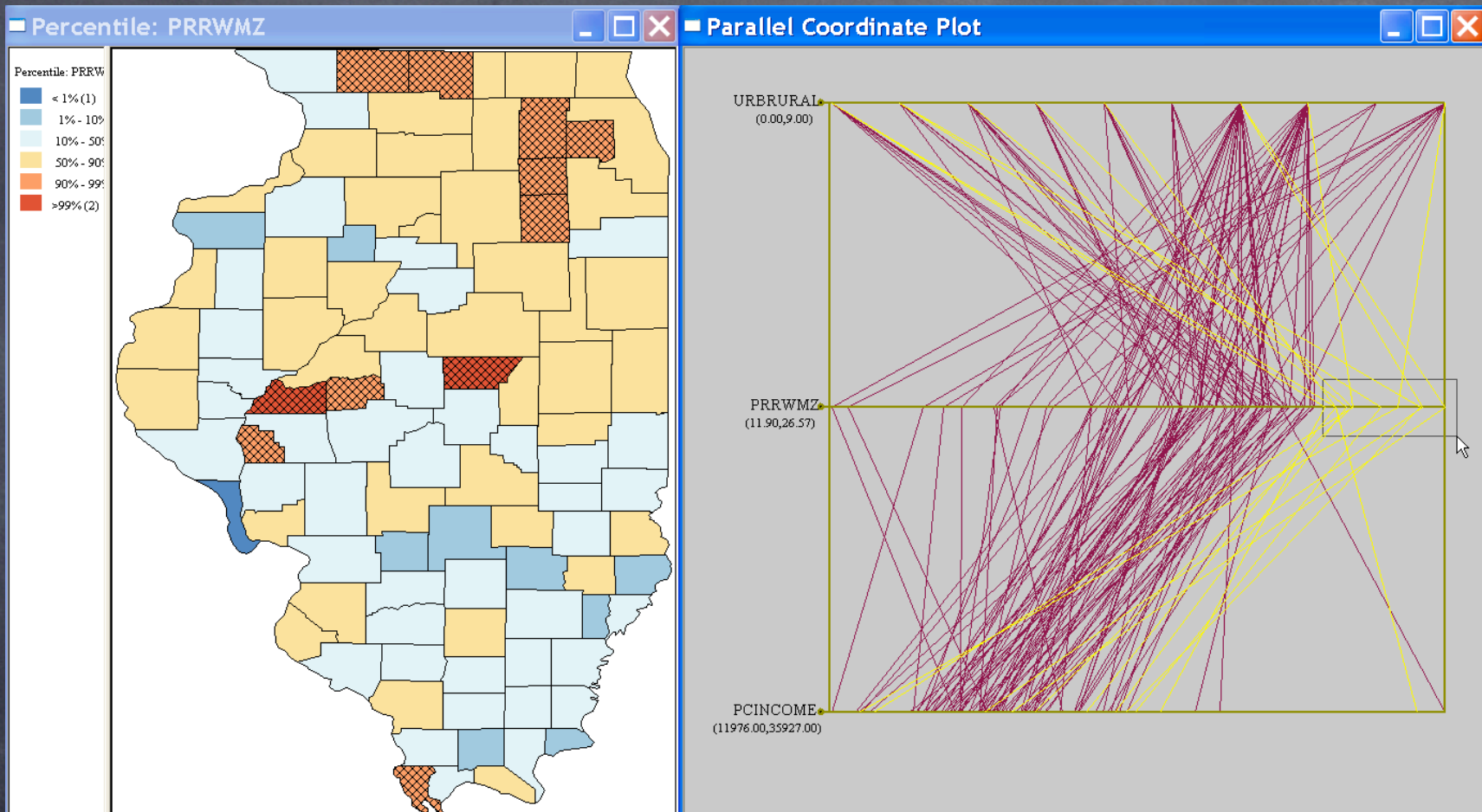
brushing the parallel coordinate plot
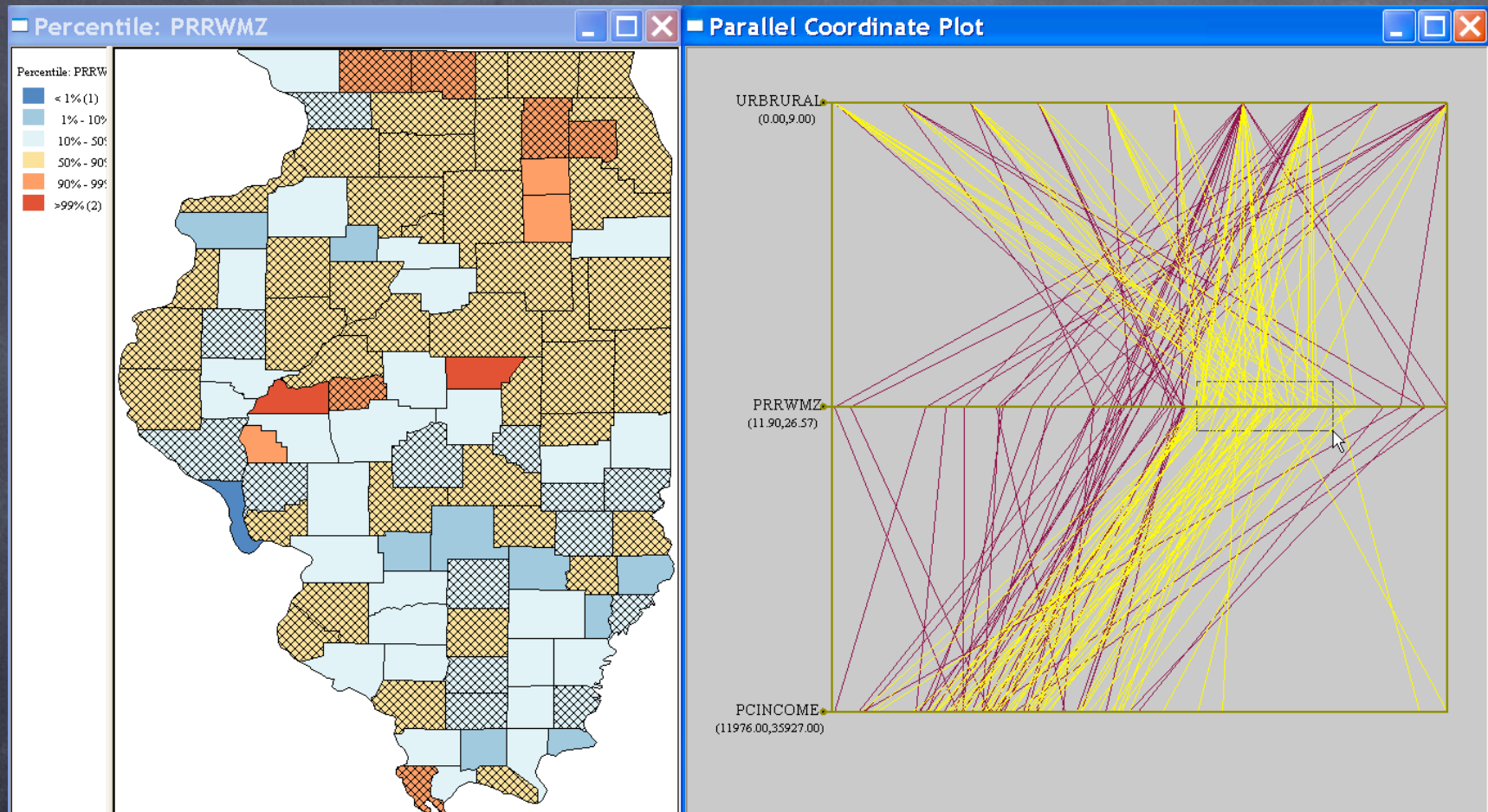
brushing the parallel coordinate plot
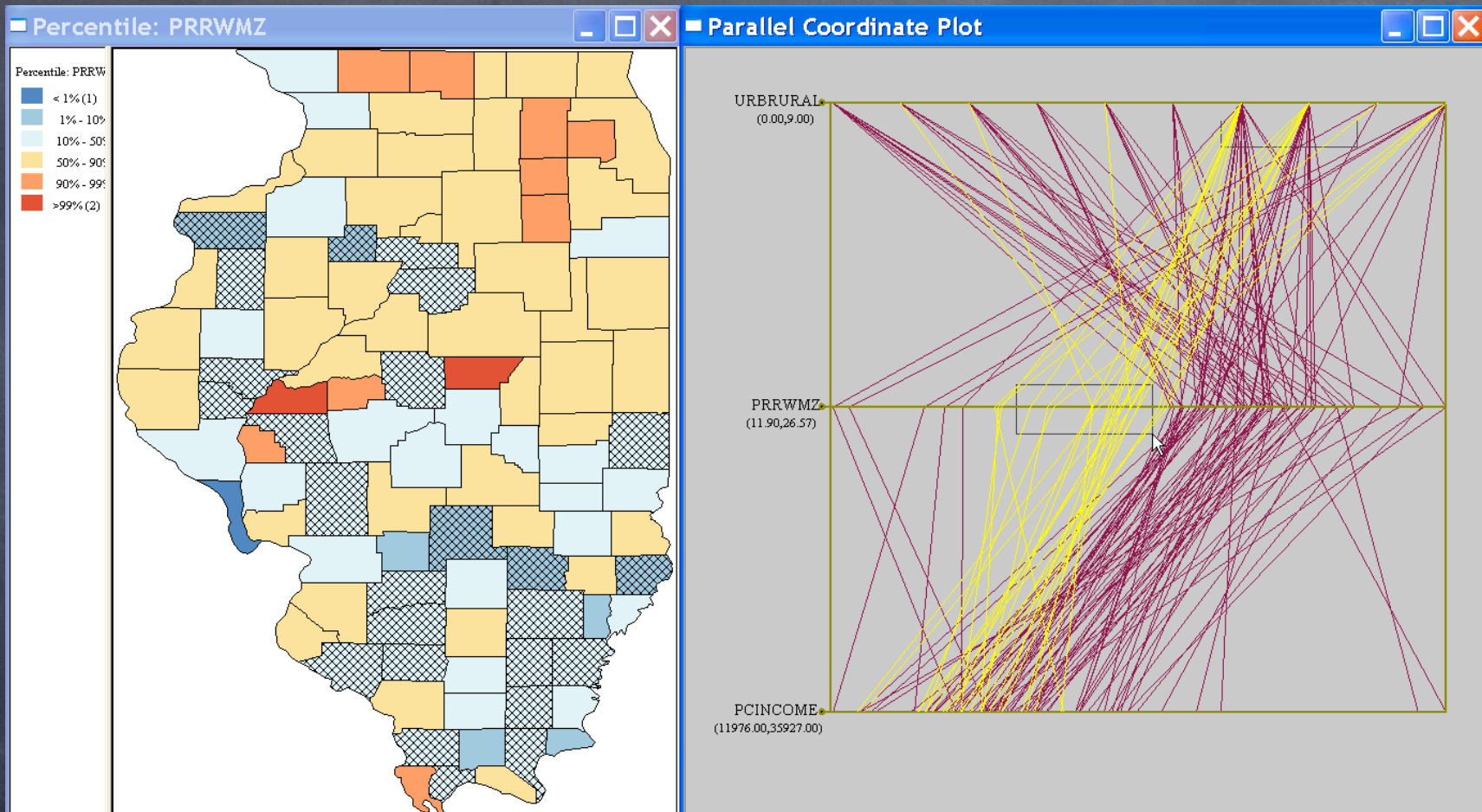
brushing the parallel coordinate plot
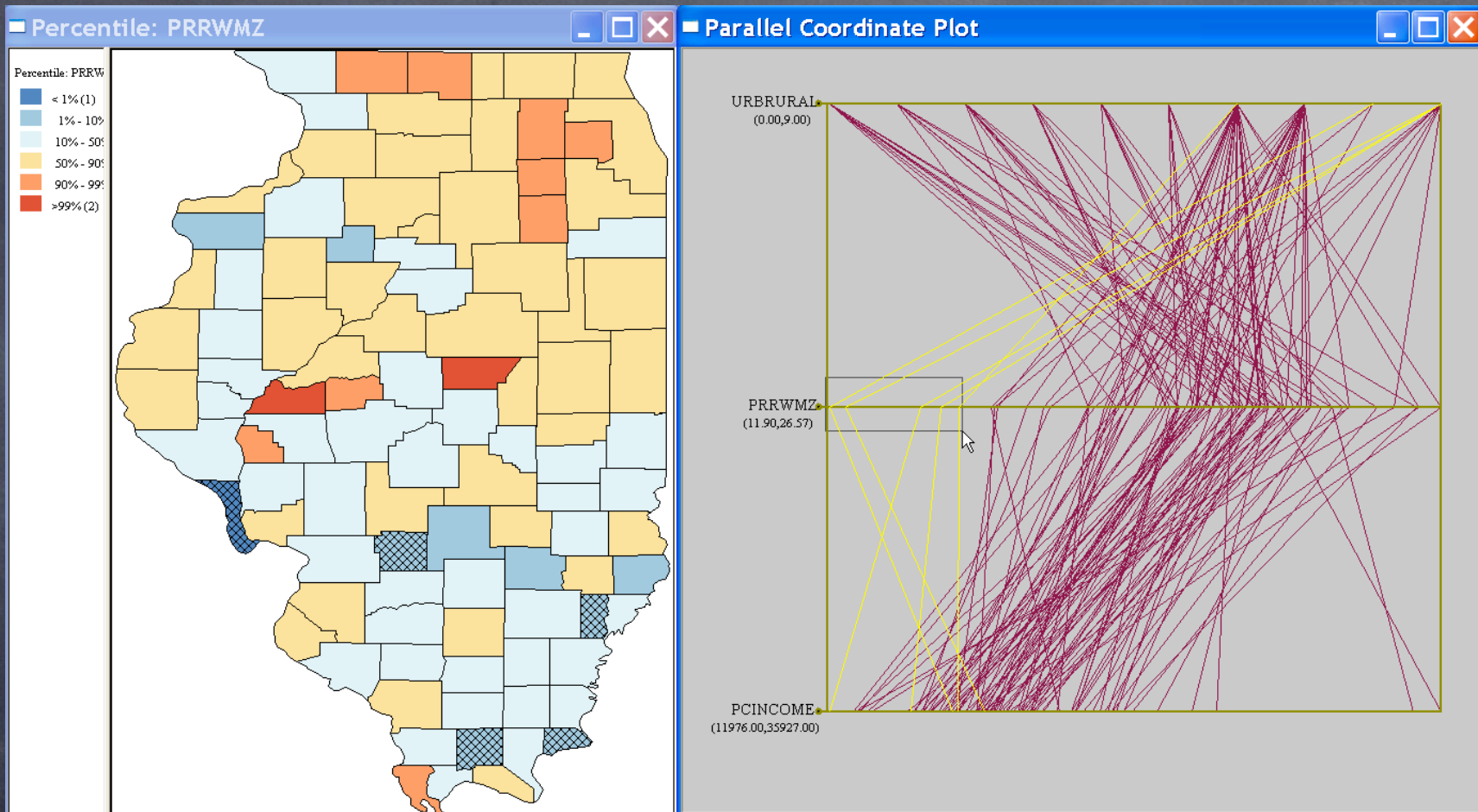
brushing the parallel coordinate plot
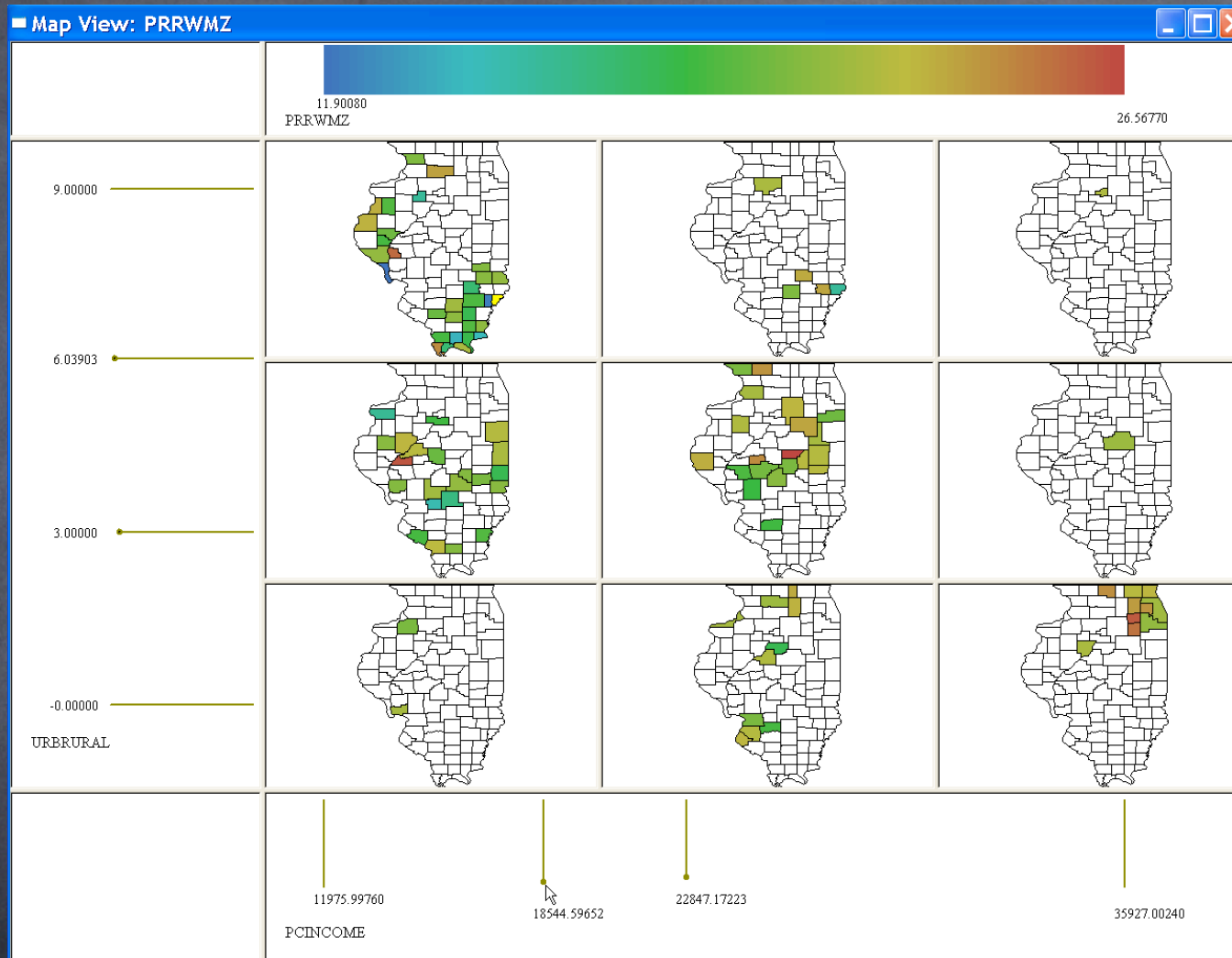
conditional maps
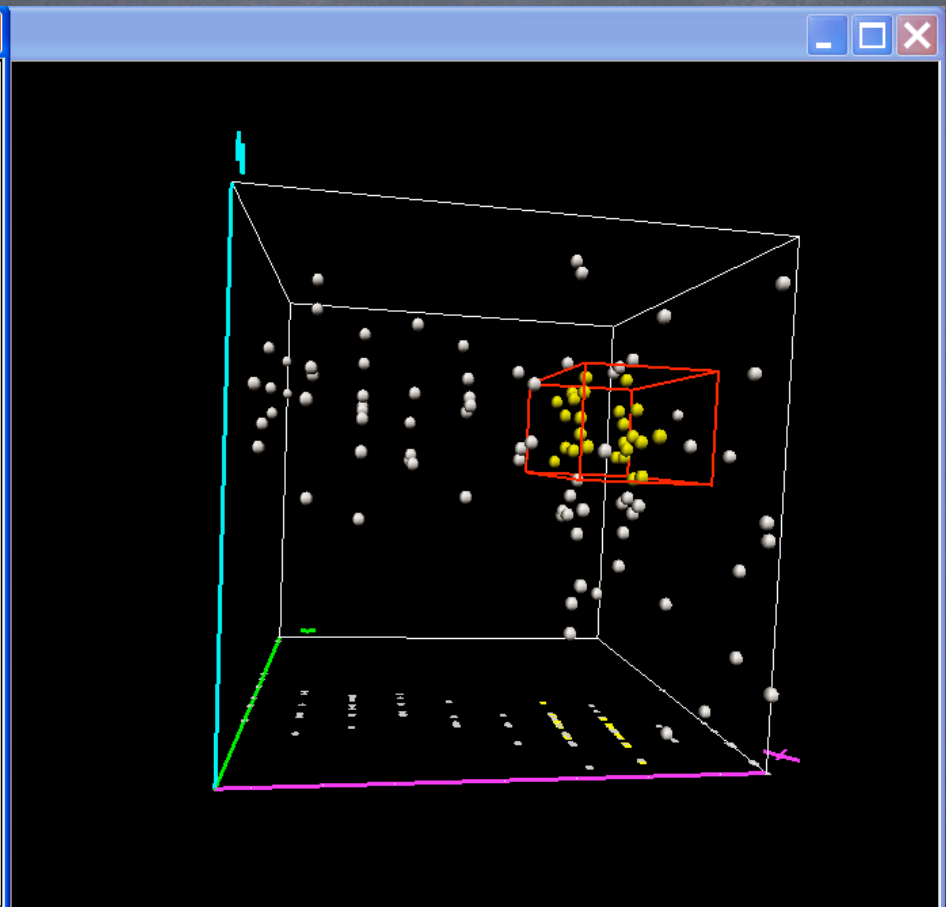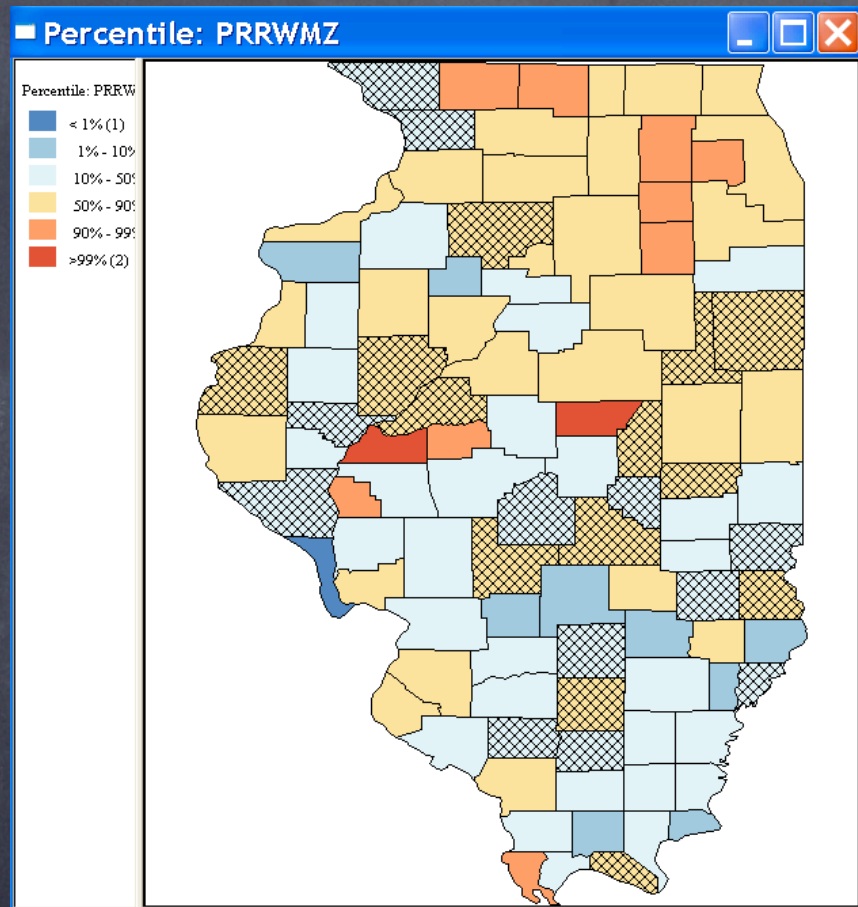
brushing the 3-D scatter plot

demonstration

Spatial Autocorrelation

- Moran Scatter Plot
  - randomization inference
  - EB adjusted Moran for rates
- Local Moran
  - LISA cluster map and significance map
  - sensitivity analysis
- Bivariate Moran Scatter Plot
- Bivariate Local Moran

Global and Local Clustering

Global Moran's I

- What is the extent of clustering in the total area?
- Is this clustering significantly different from a random spatial distribution?

Local Moran's I

- Do local clusters (high-high or low-low) or local spatial outliers (high-low or low-high) exist?
- Are these local clusters and spatial outliers statistically significant?

Global Clustering

Box Map

1990 Homicide Rates

Global Moran's I

$$I = \left(\frac{N}{S_0}\right) \frac{\sum_i \sum_j w_{ij} z_i z_j}{\sum_i z_i^2}$$

$$z_i = x_i - \mu$$

$$z_j = x_j - \mu$$

$$S_0 = \sum_i \sum_j w_{ij}$$

BoxMap (Hinge=1.5) : HR90

Randomization

permutation: 999
p-value   : 0.0010

Close

Run

I:0.3975 E[I]:-0.0003 Mean:0.1045 Sd:0.0041

Moran's I with EB Rate Standardization: HC90...

Moran's I= 0.3975   I = 0.2478

W RATES

RATES
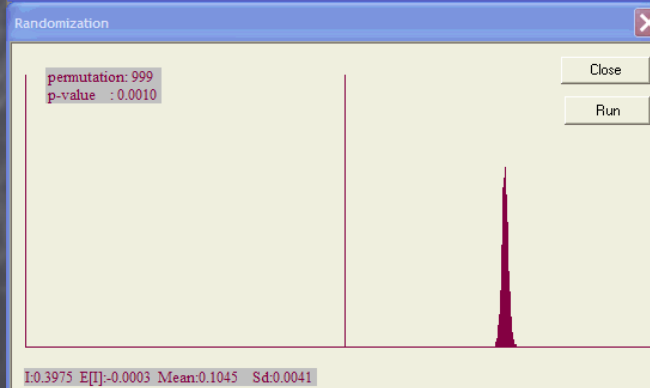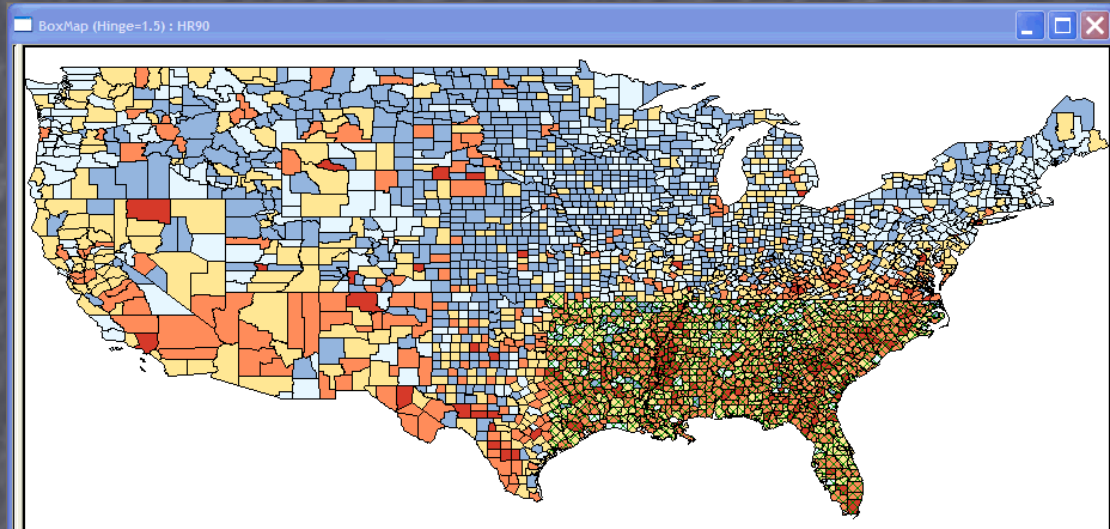
Randomization Test
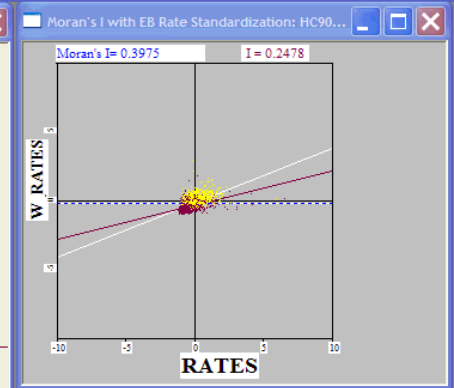
Moran Scatter Plot

Local Cluster & Significance Maps

Cluster map

Significance map
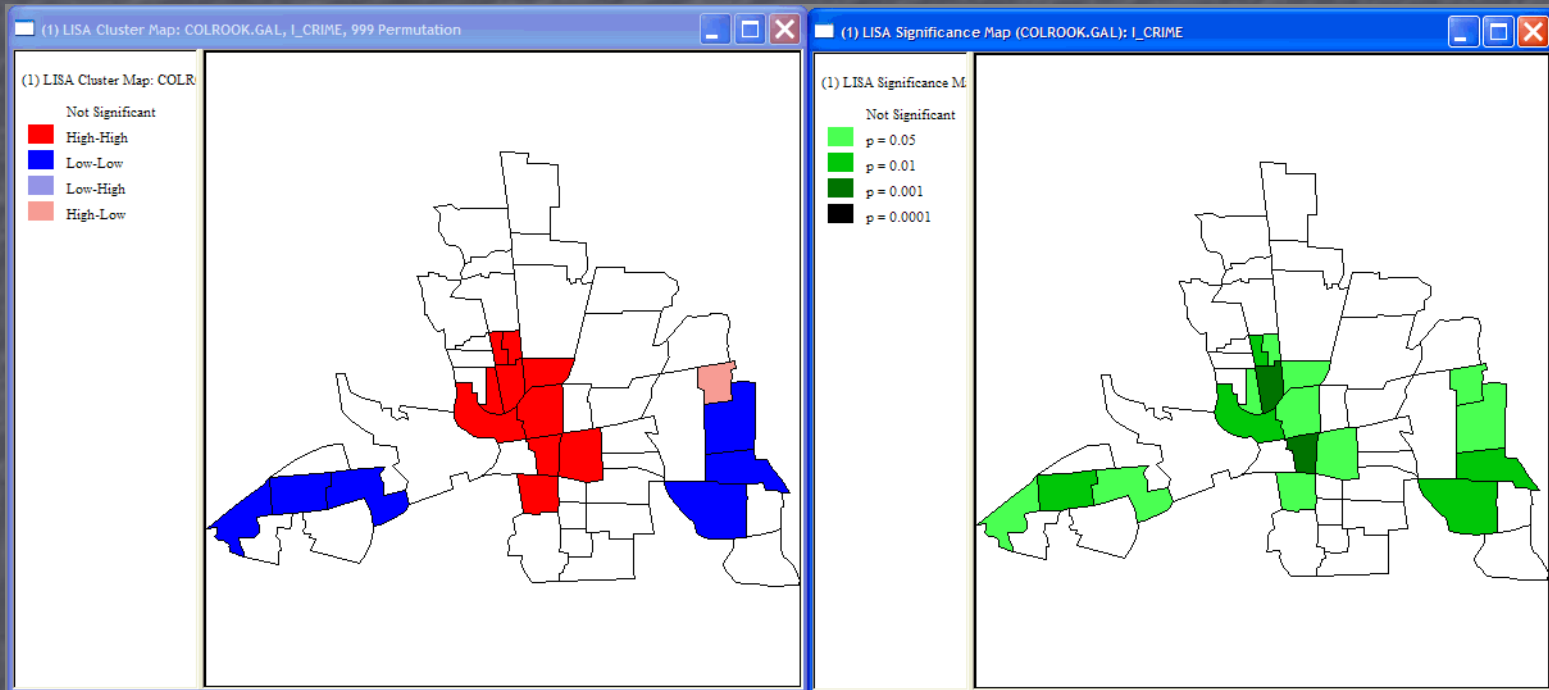
Local Indicators of Spatial Association (LISA) maps can be used to identify spatial clusters and outliers at different significance levels

Computing Local Clusters

Illustration: Mechanics behind computing hi-hi crime cluster
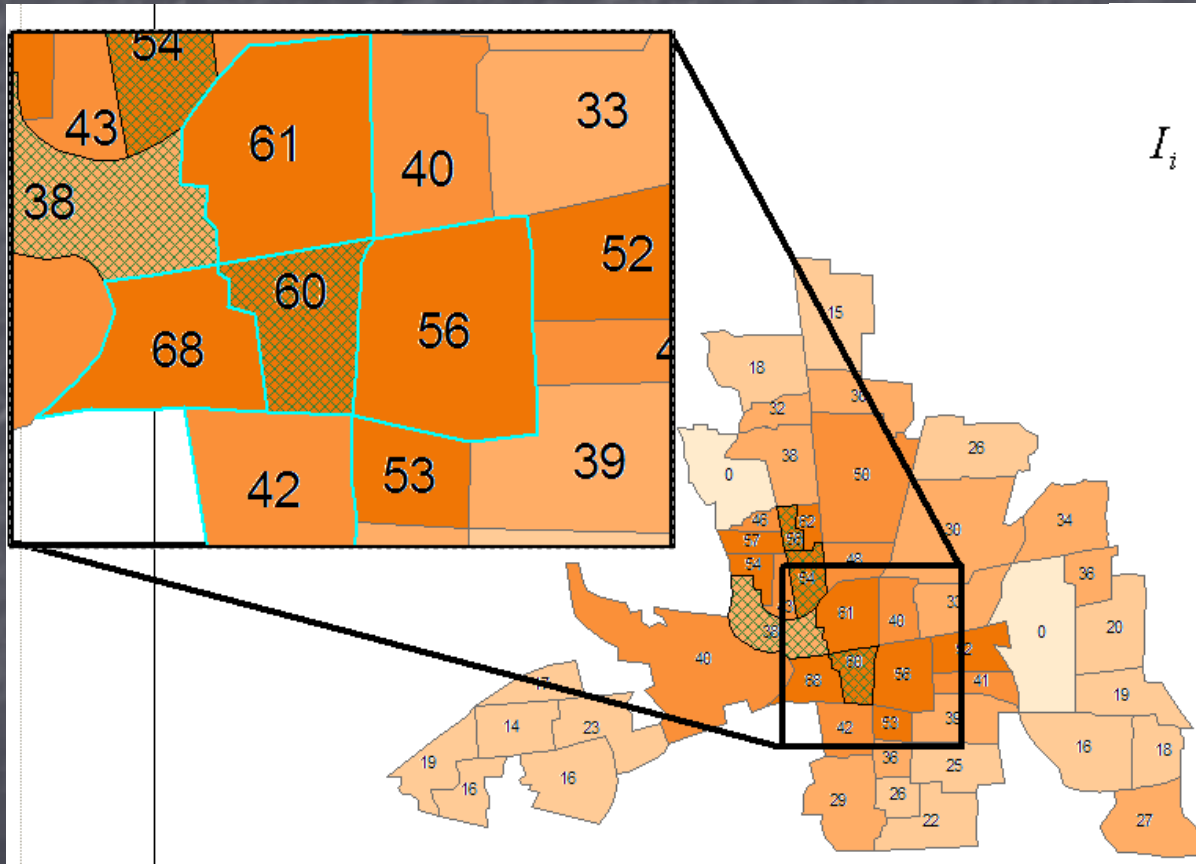
$$I_i = \frac{(x_i - \mu_x)\sum\limits_{j} w_{ij}(x_j - \mu_x)}{\sum\limits_{i}(x_i - \mu_x)^2/n}$$

Cross-product of standardized value for area i and average standardized values of neighbors j, e.g.:

Numerator (mean=57): (60-57)*[(56-57)+(61-57)+(68-57)+(42-57)]

Denominator: Sum of squared standardized values for each area i in total study area, divided by N

GeoDa uses row-standardized weights (rows sum to one)
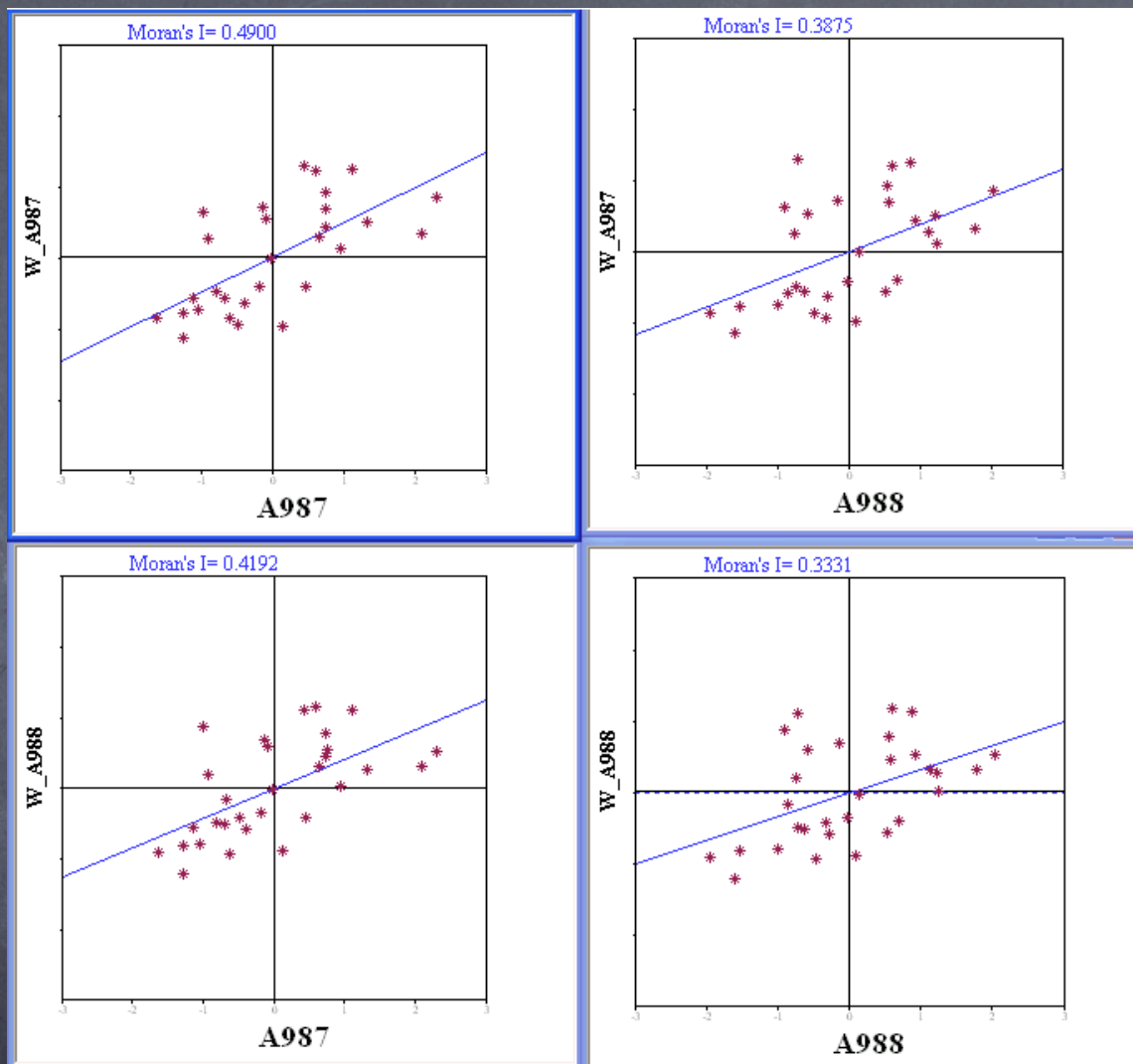
Bivariate Moran scatter plot matrix

demonstration

Spatial Regression

- OLS with Diagnostics
  - normality, heteroskedasticity
  - Moran's I
  - Lagrange Multiplier tests
- Maximum Likelihood Spatial Lag
  - asymptotic variance matrix
- Maximum Likelihood Spatial Error
  - asymptotic variance matrix
- Predicted Value and Residual Maps

OLS with Diagnostics

```
REGRESSION
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Data set             : COLUMBUS
Dependent Variable   :        CRIME   Number of Observations:    49
Mean dependent var   :     35.1288    Number of Variables   :     3
S.D. dependent var   :     16.5605    Degrees of Freedom    :    46

R-squared            :     0.552404   F-statistic           :      28.3856
Adjusted R-squared   :     0.532943   Prob(F-statistic)     :9.34074e-009
Sum squared residual:      6014.89    Log likelihood        :     -187.377
Sigma-square         :     130.759    Akaike info criterion :      380.754
S.E. of regression   :      11.435    Schwarz criterion     :      386.43
Sigma-square ML      :     122.753
S.E of regression ML:      11.0794
-----------------------------------------------------------------------
     Variable      Coefficient     Std.Error     t-Statistic    Probability
-----------------------------------------------------------------------
     CONSTANT      68.61896         4.735486       14.49037      0.0000000
         INC       -1.597311        0.3341308      -4.780496     0.0000183
        HOVAL      -0.2739315       0.1031987      -2.654409     0.0108745
-----------------------------------------------------------------------


REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER   6.541828
TEST ON NORMALITY OF ERRORS
TEST                  DF             VALUE          PROB
Jarque-Bera           2             1.835753       0.3993663

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                  DF             VALUE          PROB
Breusch-Pagan test    2             7.900442       0.0192505
Koenker-Bassett test  2             5.694088       0.0580156
SPECIFICATION ROBUST TEST
TEST                  DF             VALUE          PROB
White                 5             19.94601       0.0012792

DIAGNOSTICS FOR SPATIAL DEPENDENCE
FOR WEIGHT MATRIX : colrook.GAL  (row-standardized weights)
TEST                          MI/DF     VALUE          PROB
Moran's I (error)             0.249862  2.9376173      0.0033076
Lagrange Multiplier (lag)     1         8.7599071      0.0030792
Robust LM (lag)               1         3.0721737      0.0796429
Lagrange Multiplier (error)   1         5.8148799      0.0158911
Robust LM (error)             1         0.1271465      0.7214092
Lagrange Multiplier (SARMA)   2         8.8870536      0.0117544
========================= END OF REPORT ===========================
```

ML Spatial Lag

ML Spatial Error

```
REGRESSION
SUMMARY OF OUTPUT: SPATIAL ERROR MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            :  COLUMBUS
Spatial Weight      :  colrook.GAL
Dependent Variable  :      CRIME   Number of Observations:    49
Mean dependent var  :   35.128824  Number of Variables   :     3
S.D. dependent var  :   16.560476  Degree of Freedom     :    46
Lag coeff. (Lambda) :    0.548474

R-squared           :    0.653718  R-squared (BUSE)      :  -
Sq. Correlation     :  -           Log likelihood        :  -181.894632
Sigma-square        :   94.967743  Akaike info criterion :    369.789
S.E of regression   :    9.74514   Schwarz criterion     :  375.464725

-----------------------------------------------------------------------
    Variable     Coefficient     Std.Error     z-value     Probability
-----------------------------------------------------------------------
     CONSTANT      60.37519        5.32507      11.33791    0.0000000
          INC     -0.9610436      0.3311456     -2.902179   0.0037059
        HOVAL     -0.3031981      0.09264126    -3.27282    0.0010649
       LAMBDA      0.548474       0.1313791      4.174744   0.0000299
-----------------------------------------------------------------------


REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                                        DF      VALUE        PROB
Breusch-Pagan test                          2      18.10319    0.0001172

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL ERROR DEPENDENCE FOR WEIGHT MATRIX : colrook.GAL
TEST                                        DF      VALUE        PROB
Likelihood Ratio Test                       1      10.96521    0.0009284
========================= END OF REPORT ===============================
```
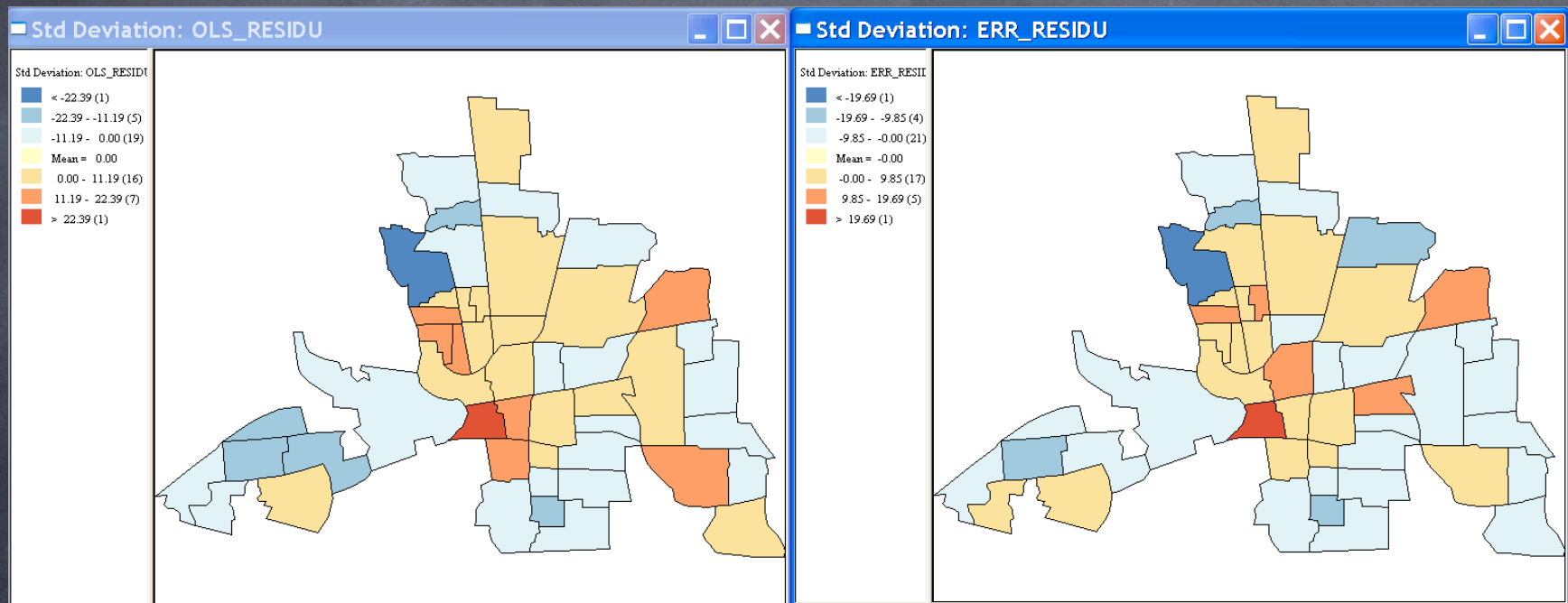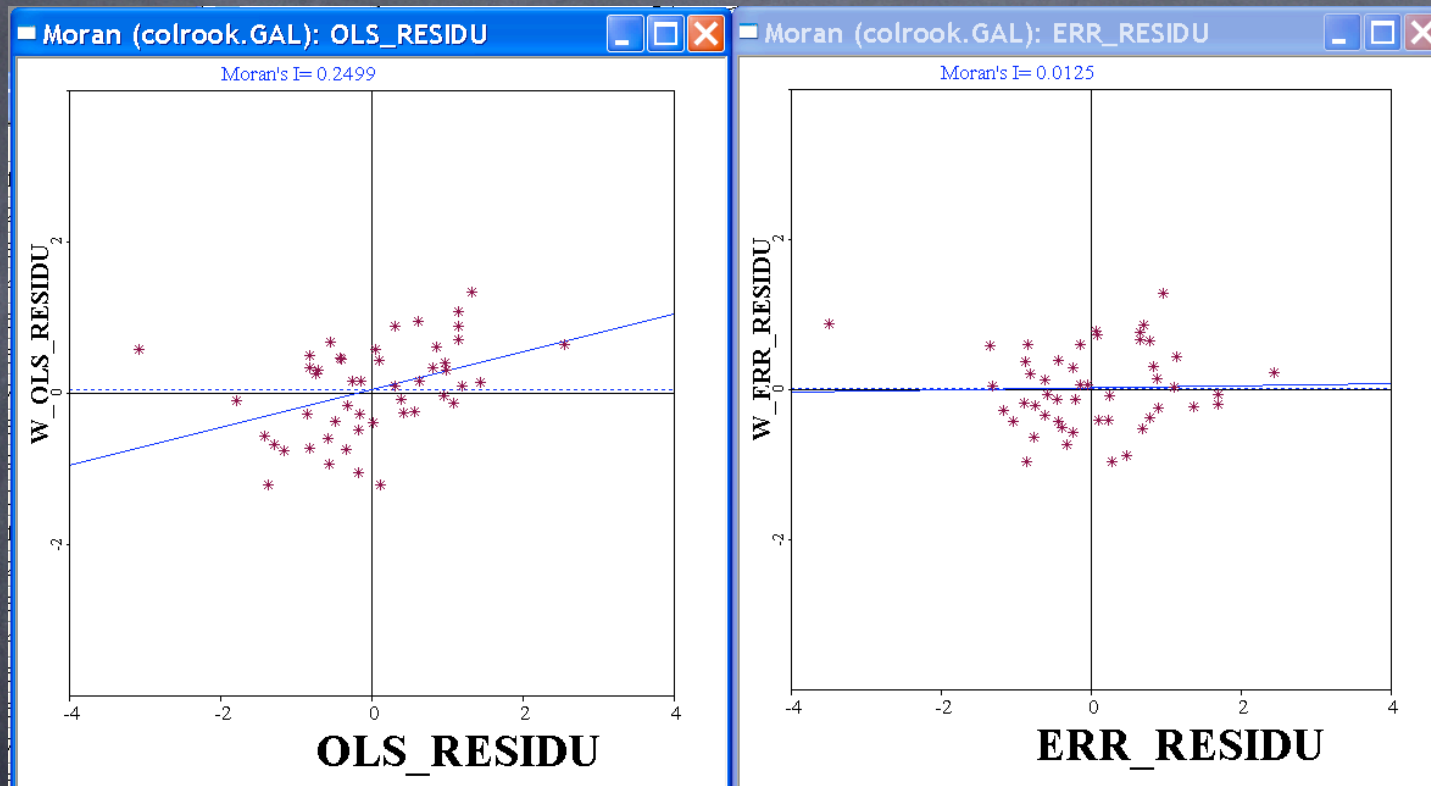
Residual Maps

OLS residuals

Spatial Error residuals

Residual Autocorrelation

Moran's I = 0.25

Moran's I = 0.01

demonstration

Help System

- Html Based
    - cross-platform
    - hyperlinks
- Resources
    - frequently asked questions (FAQ)
    - troubleshooting and bugs
    - tips
    - index of keywords
    - glossary of terms
    - references

Help System for GeoDa 0.95i

Welcome to GeoDa's Help system, which you can access by searching the entire system or by selecting links under Resources or the How-To topics below. GeoDa's two user's guides contain additional information. A search function of the system is under development.

If your question is not addressed, please post it on the Openspace mailing list. If you have suggestions for improving GeoDa's help system, please send us an email.

Most of the links in the help system are internal. However, several links access resources on the internet (such as links to *Openspace* threads). If you are using the help system without an internet connection, you will receive *"Page not Found"* error messages in these instances.

Resources

- Frequently Asked Questions (FAQ)
- Troubleshooting and Bugs
- Tips
- Index of Keywords
- Glossary of Terms
- References (including to GeoDa's Algorithms)

How to Use GeoDa's Features

Getting Started: Input, Mapping and Table Operations

- Getting Started with GeoDa
  - Load shape file - menu options -
- Creating Choropleth Maps
  - Quantile - Percentile - Box - Std Dev Maps - Selection -
- Basic Table Operations
  - Select - sort - query - new fields -
- Creating a Point Shape File
  - Input formats - import as point shape file -
- Creating Polygon Shape Files
  - Polygon from boundary file and grid - join data -
- Transforming Point and Polygon Shape Files
  - Polygons to central points and text - add centroids - points to Thiessen polygons -

Creating Choropleth Maps

- [Make a simple choropleth map](#)
- [Select items in the map](#)
- [Change the selection tool and color](#)

Make a simple choropleth map

1. After a [shape file](#) is loaded, you can adjust the width of the legend by dragging the separator between the two panes.
2. To choose a map type, right-click on the map and go to [Choropleth Map](#).
3. There are four map options:
   - [Quantile Map](#) (or)
   - [Percentile Map](#) (or)
   - [Box Map](#) with hinge 1.5 (or) and Box Map with hinge 3 (or)
   - [Standard Deviation Map](#) (or).
4. Select a variable.
5. Optionally, specify the selected variable as the default for future operations (undo the default selection at **Edit>Select Variables**).
6. For the quantile map, specify the number of categories (maximum 9).
7. To create another map based on the same shape file, click on .
8. To **zoom in, zoom out**, or go back to the **full extent** of the map, right-click and go to **Zoom**.
9. To change the color of the map, right-click and go to **Color>Map**.
10. To return to the default version of the map as it appeared at start-up, go to **Map>Reset**.
11. To export the map, go to **Edit>Copy to Clipboard** or go to **File>Export>Capture to File** (the legend needs to be copied/saved separately) or take a [screenshot](#) of the map.

[Troublehooting & Tips](#)

1. [Problems with quantile map categories](#).
2. [Taking a screenshot of your GeoDa output](#).

Index

This page contains an index of the key words and terms used in GeoDa, listed in alphabetical order.

Most of the links in the help system are internal. However, several links access resources on the internet (such as links to *Openspace* threads). If you are using the help system without an internet connection, you will receive *"Page not Found"* error messages in these instances.

Glossary of Terms

This page contains definitions of the key terms referred to in the help system, listed in alphabetical order. Where applicable, it also provides links to:

- how-to pages ("HowTo" icon),
- troubleshooting ("Trouble" icon),
- tips
- references ("Books" icon), and
- references containing the formulas implemented in GeoDa ("Formula" icon).

Most of the links in the help system are internal. However, several links access resources on the internet. If you are using the help system without an internet connection, you will receive *"Page not Found"* error messages in these instances.

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

A

Autocorrelation

See Spatial Autocorrelation.

`[i]`

B

Base Variables

See Event and Base Variables.

Box Map

Since box maps are based on the same methodology as box plots, they can be used to detect outliers in a stricter sense than is possible with percentile maps. Box maps group values such as counts or rates into six fixed categories: Four quartiles (1-25%, 25-50%, 50-75%, and 75-100%) plus two outlier categories at the low and high end of the distribution.

Values are classified as outliers if they are 1. 5 times higher than the interquartile range (IQR). IQR is the difference between the 75th percentile (Q3) and the 25th percentile (Q1) or Q3-Q1. It describes the range of the middle of the distribution since 25% of values are above the interquartile range and 25% below it.

`HowTo`

demonstration

What Next
Cross-Platform OpenGeoDa

# Motivation

- Growing demand for open source tools
  - third world
  - education, non-profit
  - open standards
- Beyond Windows limitations
  - remove constraints embedded in MS Windows
  - allow port to supercomputing (unix/linux) etc.
 - exploit MacOSX graphics
- Run GeoDa on my Mac

Approach

- Remove MFC
    - use cross-platform GUI classes, wxWidgets
- Replace MapObjects
    - recreate mapping functionality without MapObjects
    - use OpenGL and other open source graphics libraries
- Make C++ code cross-platform
    - remove dependencies on MS VC++
    - create make files to handle platform-specific issues
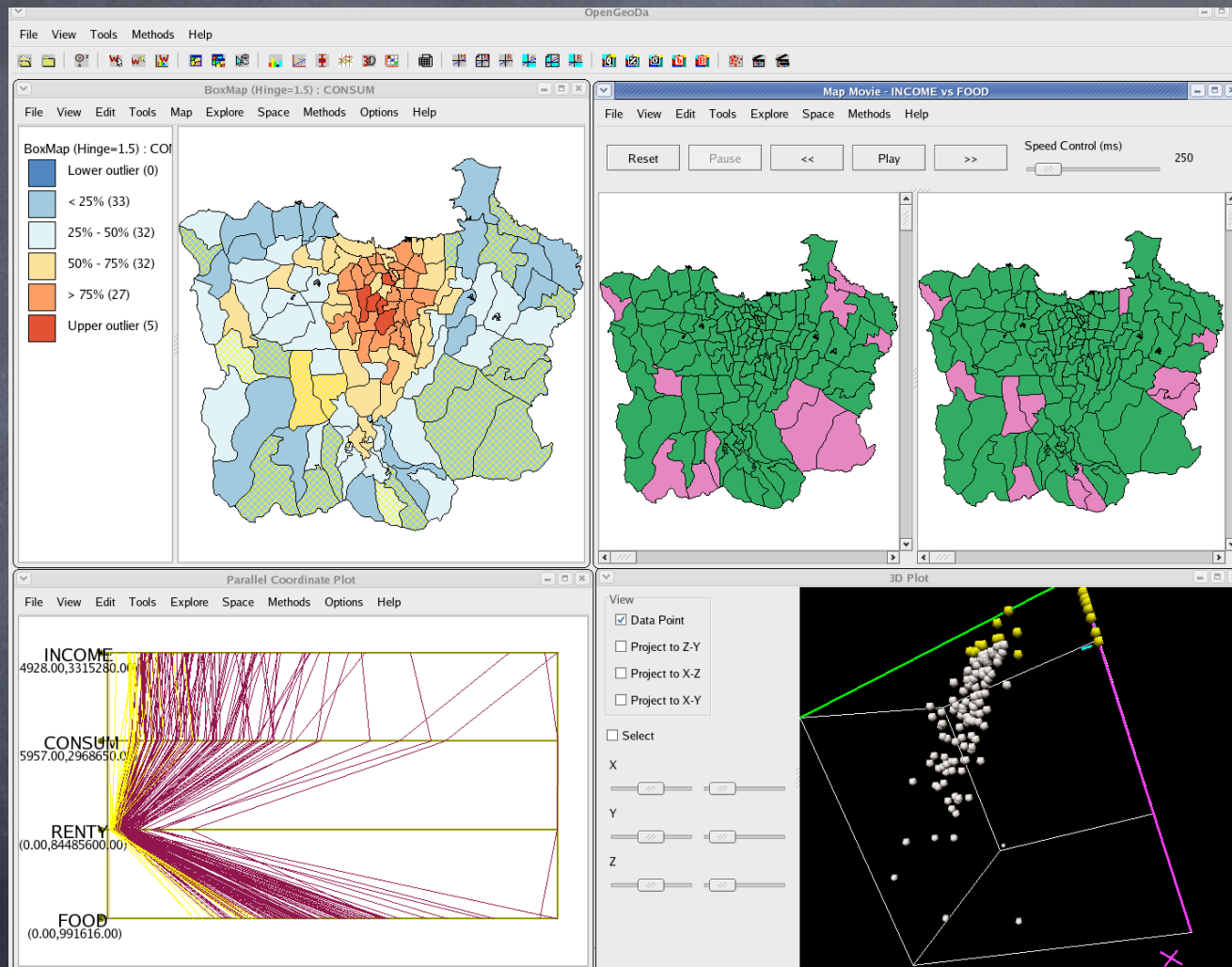- Not as simple as it sounds ...

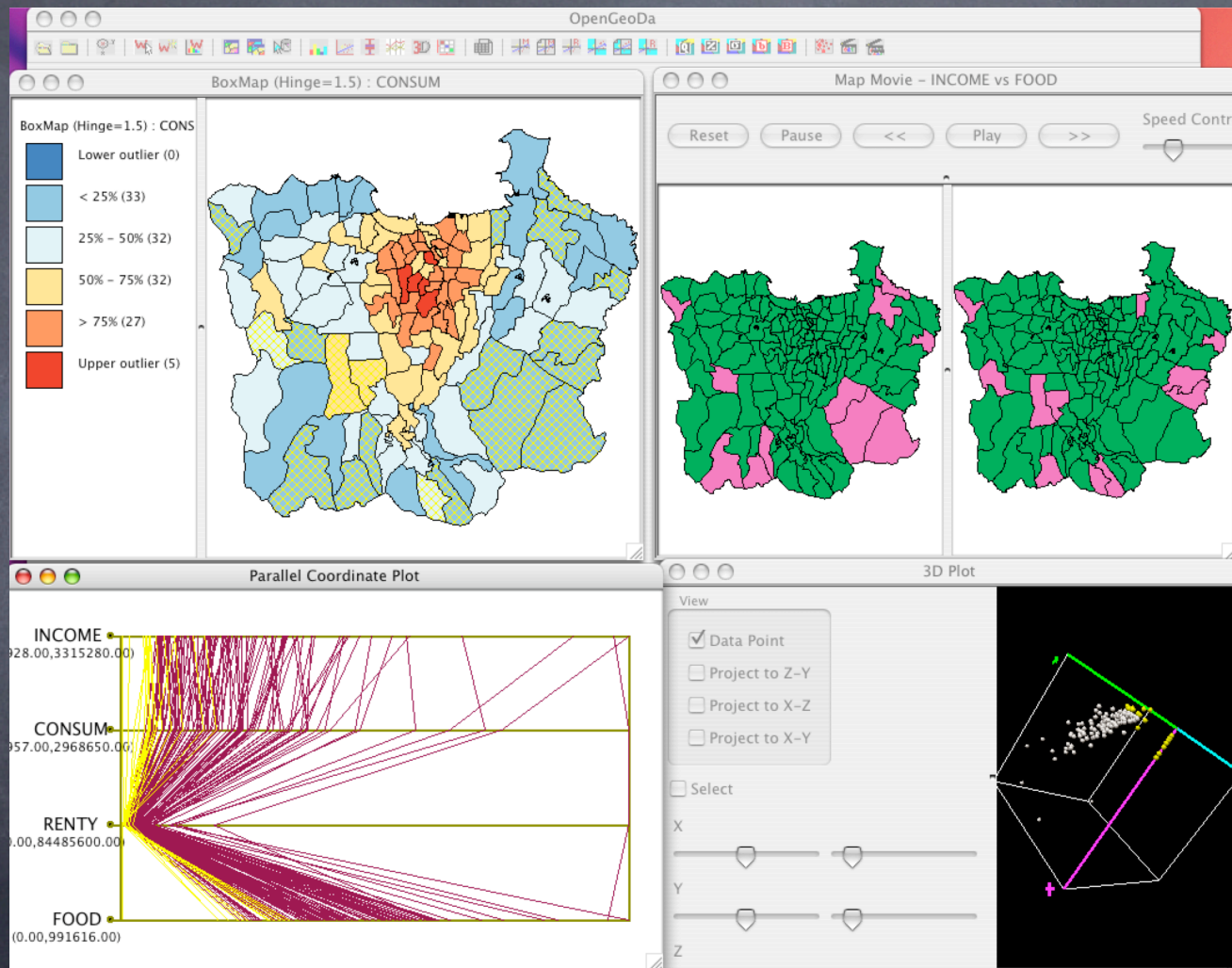# Rates, Outlier Maps, Cartogram
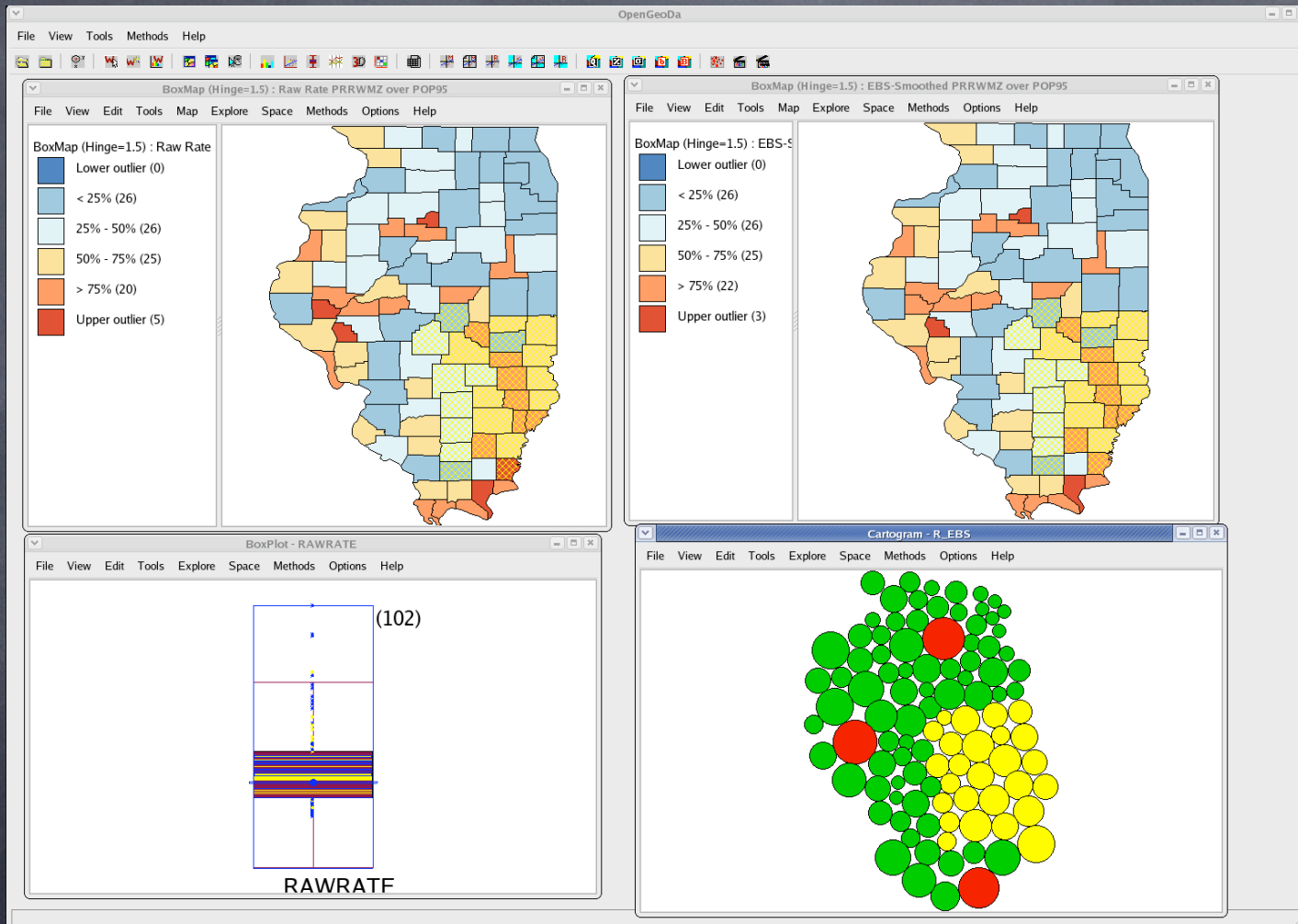
Linux

Rates, Outlier Maps, Cartogram
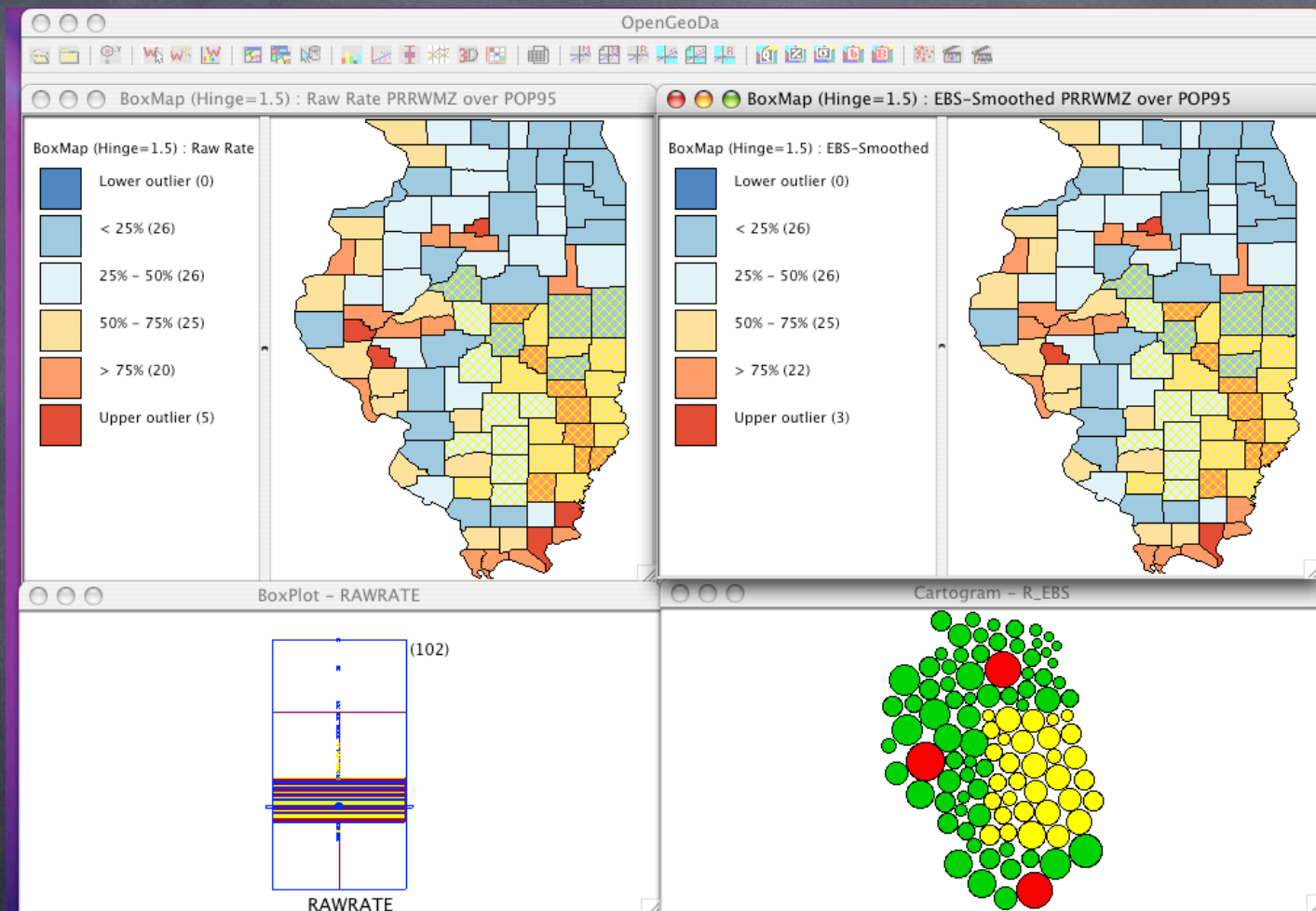
MacOS X

Multivariate ESDA

Linux

Multivariate ESDA

MacOS X

The Future

- GeoDa is here to stay
  - growing user base
- Completing cross-platform port
  - open source project
- Adding new functionality
  - generalized spatial weights
  - bivariate maps
  - IV and GM estimation
  - (enter favorite technique here) ...

Resources

- http://geoda.uiuc.edu

- Anselin (2005) Exploring Spatial Data with GeoDa: A Workbook (226pp)

- Anselin, Syabri, Kho (2005) GeoDa, an introduction to spatial data analysis. Geographical Analysis (forthcoming)